A Physics-Constrained Neural Differential Equation Framework for Data-Driven Snowpack Simulation

ANDREW CHARBONNEAU,[a] KATHERINE DECK,[a] AND TAPIO SCHNEIDER[a]

[a] *California Institute of Technology, Pasadena, California*

(Manuscript received 6 May 2024, in final form 15 March 2025, accepted 31 March 2025)

ABSTRACT: This paper presents a physics-constrained neural differential equation framework for parameterization and employs it to model the time evolution of seasonal snow depth given hydrometeorological forcings. When trained on data from multiple SNOTEL sites, the parameterization predicts daily snow depth with under 9% median error and Nash–Sutcliffe efficiencies over 0.94 across a wide variety of snow climates. The parameterization also generalizes to new sites not seen during training, which is not often true for calibrated snow models. Requiring the parameterization to predict snow water equivalent in addition to snow depth only increases the error to ~12%. The structure of the approach guarantees the satisfaction of physical constraints, enables these constraints during model training, and allows modeling at different temporal resolutions without additional retraining of the parameterization. These benefits hold potential in climate modeling and could extend to other dynamical systems with physical constraints.

KEYWORDS: Snowpack; Climate prediction; Seasonal forecasting; Neural networks; Parameterization; Deep learning

1. Introduction

Seasonal snowpacks help regulate Earth's energy balance, provide freshwater storage, and are crucial for understanding Earth's climate. They hold economic and ecological significance, supplying a majority of the western United States' (and a sixth of the world's) water supply and influencing agriculture, flood, drought, and avalanche hazards (De Michele et al. 2013; Gao et al. 2021). Their seasonal importance and susceptibility to climate change emphasize the need for ongoing modeling and monitoring on both seasonal and multidecadal time scales.

Modeling the evolution of seasonal snow for regional or global climate applications offers a challenging problem of scales; it is the bulk properties of the snow (albedo, snow-cover fraction, snow temperature, and snow water content) that are critical, yet microphysical and location-specific processes control these properties and must be taken into account. The most detailed models represent vertically resolved snowpacks, including liquid percolation, phase changes, metamorphism effects, and other types of compaction; they are often calibrated and used on the site level (e.g., De Michele et al. 2013). Models used in global climate simulations range in complexity from single-layer/bulk models to multilayer models with parameterizations for bulk properties that are calibrated with observational data (Menard et al. 2021); the horizontal resolution of these models is typically ~10–100 km where microscopic processes cannot be tractably resolved. While the laws of physics ultimately govern the evolution of snowpacks, uncertainty in how to relate essential but unresolved small-scale processes to snowpack bulk properties on the spatial scales of global models makes developing snow models a challenging task (Kapnick et al. 2018; Bair et al. 2018). This challenge is exacerbated by data availability (Menard et al. 2021; Kouki et al. 2022).

Among bulk variables in global snow models, the snow water equivalent (SWE) (m) represents the total water storage in snow. It relates to the snow depth z (m) through the bulk snow density ρ_{snow} (kg m^{-3}) and the density of liquid water ρ_{water} (1000 kg m^{-3}) as

$$\text{SWE} = (\rho_{\text{snow}}/\rho_{\text{water}})z. \tag{1}$$

It is typically used as a prognostic variable in bulk models for determining snowpack mass balance. Density and depth are critical in climate models for determining the snowpack energy balance, as they influence thermal, mechanical, and optical properties, impacting mass/energy fluxes, water retention, and spring thaw (Kouki et al. 2022; Bormann et al. 2013). However, for a given SWE, the snow depth and bulk density can vary considerably over time at a single location, or between locations under similar forcings, due to compaction, melt/refreeze cycles, and changes in the density of falling snow. These variations give rise to ongoing challenges in snow modeling.

Many prevalent snow models rely on the seminal parameterizations of Kojima (1967) and Anderson (1976) to derive snow density and depth from modeled SWE. These works modeled compaction and microstructure metamorphism by assuming a linear relationship between the strain rate of compaction and the weight of the overlying snow, suggesting empirical parameterizations that were calibrated from a select number of observational sites and laboratory experiments and analytically extended. Models such as Snow17, iSnobal, Noah/Noah-MP, CLM5, and HTESSEL (Anderson 2006; Marks et al. 2018; Niu et al. 2011; Lawrence et al. 2019; Dutra et al. 2010; Menard et al. 2021) all employ this simplified formulation based on small-scale representations for snow evolution; thus, it is integral to most current U.S. and European global snow predictions.

However, these parameterizations were primarily developed for the hydrological community instead of Earth system

Corresponding author: Andrew Charbonneau, acharbon@caltech.edu

DOI: 10.1175/AIES-D-24-0040.1 e240040

models, and their calibration to prioritize accurate SWE and subsequent local runoff induces a trade-off in density/depth errors. Such errors impact the snowpack energy balance, which is as crucial as mass balance in global climate simulations (Diro and Sushama 2018; Xu and Dirmeyer 2011), and snow depth errors remain problematic in climate predictions (Menard et al. 2021). The inadequacy for energy tracking in these formulations has been recognized for decades and spurred the development of more detailed and realistic alternatives such as CROCUS or SNOWPACK, which evolve the metamorphism of snow microstructure (Vionnet et al. 2012; Lehning et al. 2002; Brun et al. 1989). While these complex models address the limitations of Anderson's and Kojima's original formulations, they require site-specific calibration and struggle with computational scalability for global applications. This necessitates more efficient representations for global models that can accurately predict snow depth/density. With the push toward even finer localized (1–10 km) land modeling, simple yet accurate models are increasingly essential for computational efficiency (Clark et al. 2015; Schär et al. 2020; Ban et al. 2021).

Advances in computing and sensing technology have led to initiatives in data assimilation and machine learning (ML) aimed at disrupting long-standing parameterizations. Today, extensive snow depth observations surpass the resolution and precision of SWE and density data/estimates (Fontrodona-Bach et al. 2023; Bruland et al. 2015). Improving snow models with these resources has become a dominant avenue of hydrology research, with a bias toward SWE modeling, like incorporating remotely sensed depth into iSnobal to infer SWE (Hedrick et al. 2018). Several ML models (e.g., Bair et al. 2018; Meloche et al. 2022; Duan et al. 2024; Steele et al. 2024) predict SWE or depth from meteorological and topographical inputs in specific regions. Such models show satisfactory snowpack estimation but frequently yield errors over 15% when tested, especially beyond their training or calibration locations (Meloche et al. 2022; Ebner et al. 2021; Viallon-Galinier et al. 2020). The ability of these empirical models to generalize to new locations or future climates and act as a universal model is limited, and their statistical or black-box nature does not inherently respect physical constraints, impeding their integration into land models (De Michele et al. 2013; Gao et al. 2021). Combining depth observations and ML techniques to improve depth/density parameterizations has the potential to lead to improved simulation of key variables, benefiting global climate modeling. Capitalizing on this opportunity demands a representation that can generalize to many snowpacks and integrate with existing large-scale models.

This work presents a novel hybrid approach to parameterization, combining physical principles with empirical modeling that structurally guarantees compliance with prescribed bounds (e.g., physical consistency or conservation conditions). We showcase its utility in designing an alternative parameterization for snow depth, with ramifications for global climate and seasonal simulations. We also create a quality-controlled dataset for snow modeling and make it publicly available. Learning physically informed representations from observational data across many locations enables robust performance that can generalize to new locations without recalibration. The customizability of the approach permits straightforward adaptation to different operational requirements and constraints, demonstrating additional capabilities with minimal adjustment. This offers a flexible, efficient, and scalable framework that is adaptable as the field evolves. The proposed approach exhibits a versatile means for enforcing (or learning) any function-based threshold on an optimizable model without modification of the training metrics, which can contribute to contemporary global snow modeling as well as other physics/ML hybrid models.

2. Methodology

a. Overview

Our model choices leverage ML for seasonal snow simulation in climate models, prioritizing generalizability and computational efficiency. Contemporary paradigms in hydrology research focus on parameterizing SWE from z and other data to constrain its value over global grids. However, within global climate models, SWE evolution is already well constrained by explicitly implemented physical laws enforcing mass conservation and relatively well-understood fluxes such as sublimation, precipitation, and melt. The subsequent conversion of simulated SWE to variables like z or ρ_{snow} is typically left to long-standing parameterizations. These parameterizations can be precalibrated offline prior to use in a snow model or are often further tuned "online" within a snow model. The data used for calibration can be indirect (non-SWE data), such as energy and water fluxes or land surface albedo, or be direct measurements of snow depth or density, or even gridded SWE estimates. The resulting global simulations are sometimes employed in refining gridded SWE for calibrating other models, creating circular estimation and biases.

Given the relative abundance of observations of snow depth z alongside additional snow variables, this instead justifies parameterizing z (from physics-constrained SWE) as an alternative to established formulations to address these limitations, so that tighter relationships can be determined and evaluated on the basis of direct, high-quality data. By using primary observational data instead of assimilated/reanalysis data, this approach can avoid biases and inaccuracies, ensuring more faithful representations of the underlying processes.

We model the rate of change in snowpack height (m s^{-1}) by an ordinary differential equation (ODE) represented by an artificial neural network M:

$$\frac{dz}{dt} = M(z, \text{SWE}, \varphi, R, \upsilon, T_{\text{air}}, P_{\text{snow}}), \tag{2}$$

where SWE is the snow water equivalent (m), φ is the relative humidity (between 0 and 1, the used data are measured with respect to liquid water), R is the broadband solar radiative energy flux (W m^{-2}), υ is the wind speed (m s^{-1}), T_{air} is the air temperature (°C), and P_{snow} is the liquid water–equivalent rate of snowfall (m s^{-1}). The 1D column approach permits application over any spatial grid. The chosen input variables only indirectly encode location and time dependencies through the environmental input variables, allowing the model to function in areas where topographical or temporal information

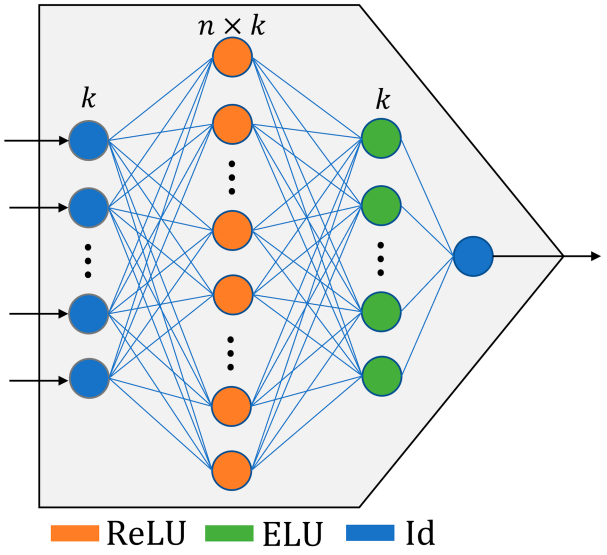

FIG. 1. Structure of the model's predictive component. Blue lines indicate a trainable linear transformation of the input (of k scalar variables), including a bias. Colors indicate the activation function used upon collection at the node (ReLU = rectified linear unit; ELU = exponential linear unit; Id = identity), as noted in the legend. The hyperparameter n sets the width of the internal mixing layer.

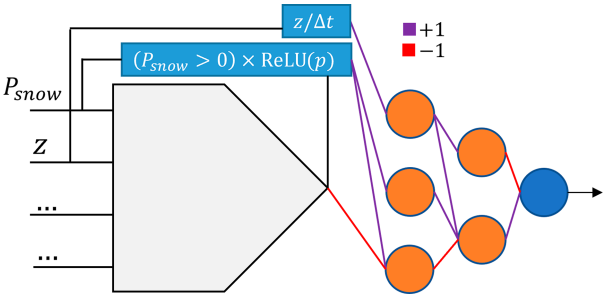

FIG. 2. Architecture of M, highlighting the constraint component attached to the predictive component (the gray pentagon) described in Fig. 1. The chosen structure enforces increasing snow depth only under precipitation and nonnegativity of snowpack height and is equivalent to a maximum/minimum block on the output. Weight colors indicate the constant's sign, and activation functions follow the color scheme given in Fig. 1.

is unavailable. This choice aims to enable learning about universal linear and nonlinear physical processes that apply independent of time, season, and location. Using a feed-forward neural network dependent only on the current system state aligns with land surface models, as it matches the differential equation format used for other variables. This model is also adaptable for different applications or when SWE is unavailable (see sections 2b and 3e).

b. Model structure

The model M consists of two components. The first is a "predictive" network with trainable weights to generate a dz/dt prediction (Fig. 1). For computational simplicity, only two hidden layers were used, which can also be interpreted as a regression on once-transformed features, with the transformational layer width set by the hyperparameter n scaled by the number of input features k (see Fig. 1). Inputs are easily exchangeable for alternative use cases or target predictions.

The second component consists of fixed-weight dense layers with rectified linear unit (ReLU) activation, designed to impose explicit ("hard") constraints on the predictive model. This approach allows for enforcing minimum/maximum thresholds on any predictive model, without introducing penalties into the calibration metrics. Although more advanced methods exist in literature or modern coding packages (Jiang et al. 2020; Dong and Ni 2021; Beucler et al. 2021), our simplistic approach offers multiple advantages. Primarily, constraints are applied throughout training, leading to better gradient updates within prescribed limits (see Table C2 for comparison). This flexible framework supports most functions or specialized constructions, including "learned" constraints, and can be scalably implemented in environments with minimal to no ML support, offering potential for many fields (for more on this process, see appendix A). By guaranteeing physical constraints, it ensures stability during time stepping and is conducive for integration into larger models without violating prescribed bounds such as conservation or consistency equations.

c. Threshold constraints for snowpack prediction

Constraints for dz/dt should keep the depth tendency within physical limits to enable generalizability and stability when M is integrated over time. This initial study selected the following basic constraints:

- Enforce depth nonnegativity within a time step Δt, i.e., $M \geq -z/\Delta t$.
- Enforce depth inability to increase without snowfall, i.e., $P_{\text{snow}} = 0 \Longrightarrow M \leq 0$. Processes like wind drift violate this constraint, but such effects are small in our data (see appendix B).

These constraints can be expressed as threshold functions, the lower as $f_- = -z/\Delta t$ and the upper as $f_+ = \text{ReLU}(p) \times \mathbf{1}_{P_{\text{snow}}>0}$, where p is the output of the predictive component and $\mathbf{1}$ is the indicator function. For these choices, f_- is nonpositive and f_+ is nonnegative, which simplifies the constraint layer structure (see appendix A), resulting in a final structure for M as depicted in Fig. 2.

The first constraint includes the time step Δt, but this does not explicitly affect the time dependency or resolution of the parameterization. The predictive component contains no time nor time step dependence. Adjusting Δt scales the constraint appropriately without altering the predictive component's output, enabling the model's use in adaptive time step schemes (rescaling the constraint per time step). This means the model in principle only requires training with data at one temporal resolution, though we would anticipate improved time step independence if training data incorporated varied time intervals, or dz/dt values in the minimum range anticipated during usage.

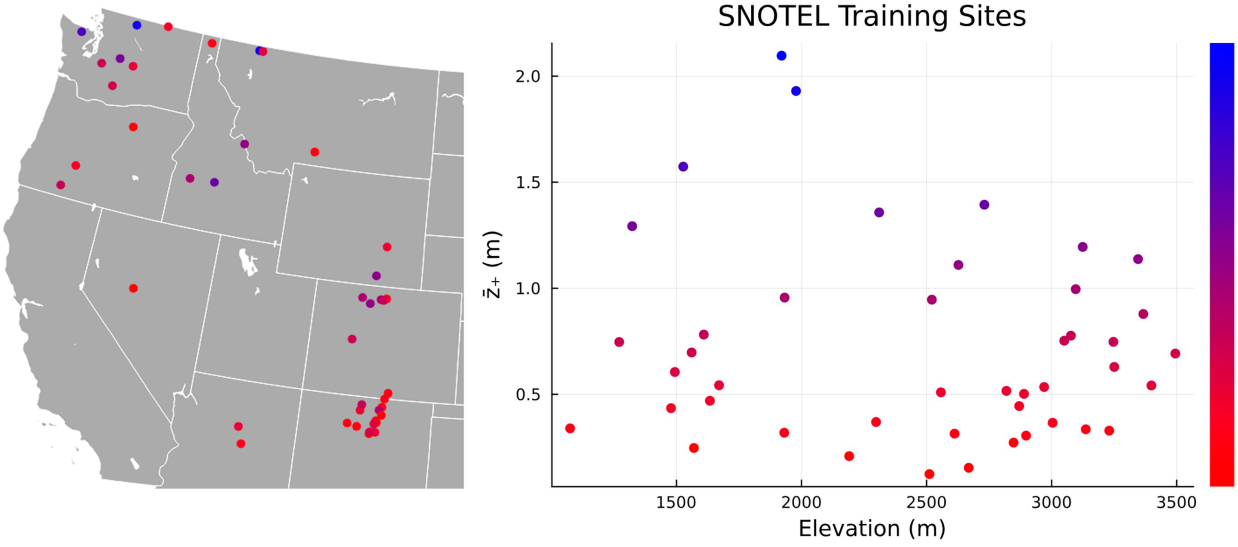

FIG. 3. Distribution of SNOTEL sites used for training the network. (left) Training sites as visualized over the United States. (right) Training sites visualized with elevation vs their average nonzero snowpack height $\bar{z}_+$. The color bar is a visual indicator of $\bar{z}_+$ for visualization on the spatial map.

d. Data

We required training data with simultaneous snow and meteorological measurements, preferring collocated primary ground observations over reanalysis data due to known discrepancies (Meyer et al. 2023). One primary source is the U.S. Snowpack Telemetry (SNOTEL) network by the Natural Resources Conservation Service (NRCS). Data from 44 SNOTEL sites in the contiguous United States (CONUS) have simultaneous availability of the necessary inputs, of which their entire reporting histories until 1 February 2024 were collected. These sites span diverse climates (see Fig. 3), enhancing generalizability. For testing, seven Alaskan SNOTEL sites had the necessary inputs and were similarly collected, plus standard evaluation data from Kühtai, Austria (Krajči et al. 2017), Col de Porte, France (Lejeune et al. 2019), and the Reynolds Mountain East catchment (Reba et al. 2011) to assess the ability of M to generalize to out-of-sample data. We additionally used data from Sodankylä, Finland (Essery et al. 2016), the upper Rofental (Warscher et al. 2024), and Yala Basecamp (Stigter et al. 2021; Shea et al. 2015) in the Himalayas to test performance across uncalibrated elevations and climate types.

SNOTEL data have known quality issues, such as underrepresenting complex mountainous terrain and underreporting precipitation (Meyer et al. 2012; Serreze et al. 1999), along with periods of biased or unphysical values (Hill et al. 2019). To ensure data suitability for model training, we applied established cleaning measures formulated by Serreze et al. (1999) and refined by Yan et al. (2018) to SNOTEL daily snowpack data. Gauge undercatch was corrected as in Livneh et al. (2014), and temperature biases were addressed following the SNOTEL correction release (Atwood et al. 2023). Since no consistent quality control procedures exist at present for SNOTEL meteorological or snow depth data beyond outlier tests by Hill et al. (2019), we developed a custom procedure (see appendix B). Snow fraction (to obtain snowfall from total precipitation) was estimated using the (T_{air}, φ) bivariate logistic model from Jennings et al. (2018), shown to have over 88% accuracy. From the cleaned data, we derived dz/dt, $d\text{SWE}/dt$, and P_{snow} for days with complete data, excluding all data with $\Delta t > 1$ day.

The training data were averaged (preserving start-of-window z and SWE) over a consecutive N-day moving window, with N left as a hyperparameter. This enabled exploration of the trade-off of spreading out discretized SNOTEL data (z to the inch and SWE to 0.1 in.) for smoother regression learning versus preserving extreme values critical for predictions. Days with unphysical values (zero SWE and nonzero z) or no snowpack were removed to eliminate uninformative zeros in the target space. Features were scaled by their standard deviations, and the target by its absolute maximum, with scaling constants fixed into M to spare user preprocessing. This resulted in 58 484 usable sensor days out of 105 636 for training and 35 618 sensor days for testing. The complete dataset and generating code are publicly available (see data availability statement below).

e. Training

Accurate prediction of extreme values is vital in snowpack modeling. Underpredicting extreme dz/dt can prevent rapid snowpack growth or depletion, lagging snow presence early in the season or maintaining snow into the summer, which skews albedo, runoff, and energy calculations. Models like Noah, CROCUS, and SNOWPACK have struggled with this challenge (Gao et al. 2021; Luijting et al. 2018; Lundy et al. 2001; Wever et al. 2015; Vionnet et al. 2019). Standard regression tends to underpredict extremes, so we used a custom loss function that can emphasize extreme values:

$$L = \frac{1}{N_d}\sum_{i=1}^{N_d} w_i |y_i - \hat{y}_i|^{n_1}, \qquad w_i = 1 + |y_i|^{n_2}. \tag{3}$$

Here, N_d is the number of batched training data, $\hat{y}_i$ and y_i are the prediction and target data, respectively, and n_1 and n_2 are constant positive real numbers. Using ($n_1 = 1$, $n_2 = 0$) or ($n_1 = 2$, $n_2 = 0$) is equivalent to optimizing the average L_1 or L_2 losses, respectively. Hyperparameter tuning followed a leave-one-out approach, using averaged and filtered data from 43 of 44 sites. Validation scores were generated over the remaining sites using unaveraged, unfiltered data, and then averaged over all sites to guide hyperparameter selection. For time stepping (see section 2f), the optimal n_1 was found with L_2 training and $n_2 > 0$, highlighting the importance of extreme points. We note that optimal hyperparameters varied between regression and time-series tasks; see appendix C for more details.

The model was implemented in Julia and the Flux framework (Innes et al. 2018; Innes 2018), with the root-mean-square propagation (RMSProp) optimizer (Hinton et al. 2012). Training for 100 epochs (100 passes over all data) takes under 30 s on one Intel i9 CPU without graphics processing unit (GPU) usage, with model storage under 3 KB. Time and memory benchmarking of the network are listed in Table C4 in appendix C.

f. Testing

Model performance was tested by time stepping the dz/dt equation with an explicit Euler method:

$$\hat{z}_{i+K} = \hat{z}_i + (K\Delta t)M(\hat{z}_i, \mathrm{SWE}_i, \varphi_i, R_i, \upsilon_i, T_{\mathrm{air},i}, P_{\mathrm{snow},i}). \quad (4)$$

The integer K specifies sequential data transitions ($K = 1$) or "gaps" to traverse in time-series data ($K > 1$) due to missing or cleaned data. The built-in constraints of M ensure nonnegative z values when the step size is the designated Δt (i.e., when $K = 1$; for providing other Δt, see section 3f), though this choice of time-stepping procedure can create negative z when $K > 1$ (rates constrained for step size Δt are applied over $K\Delta t > \Delta t$). In such cases, negative z values were set to zero, and similarly, time series were "reset" to observed values $\hat{z}_{i+K} = z_{i+K}$ whenever $K > K_{\max} = 5$ days, to avoid attributing error from the method choice to M for fair evaluation. When M is used within a global model simulation, such gaps would not occur since the inputs would be available at every step, allowing all z_i to obey prescribed bounds. Selecting $K_{\max} > 1$ day additionally reduced the number of resets that would otherwise beneficially skew performance metrics, with $K_{\max} = 5$ ensuring resets in less than 2.3% of cases (median frequency 0.19%), with many over whole years or no-snow periods (not impacting snow simulation). Evaluation metrics included root-mean-squared error (RMSE), mean absolute error (MAE), bias (B), and median percent error (MPE; on dz/dt for regression and z for time series). For time series, we included the Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe 1970) and snowpack percent error (SPE) MAE/$\overline{z}_+$, where $\overline{z}_+$ is the mean nonzero snow depth.

By swapping SWE and z features, the network can also be trained to predict $d\mathrm{SWE}/dt$ from z, allowing SWE to be simulated from depth data. This enables standalone modeling using only weather inputs, with two networks $\tilde{M}_z$ and $\tilde{M}_{\mathrm{SWE}}$, separately trained to run in a coupled fashion:

$$\widehat{\mathrm{SWE}}_{i+K} = \widehat{\mathrm{SWE}}_i + (K\Delta t)\tilde{M}_{\mathrm{SWE}}(\hat{z}_i, \widehat{\mathrm{SWE}}_i, \varphi_i, R_i, \upsilon_i, T_{\mathrm{air},i}, P_{\mathrm{snow},i}), \quad (5)$$

and

$$\hat{z}_{i+K} = \hat{z}_i + (K\Delta t)\tilde{M}_z(\hat{z}_i, \widehat{\mathrm{SWE}}_i, \varphi_i, R_i, \upsilon_i, T_{\mathrm{air},i}, P_{\mathrm{snow},i}). \quad (6)$$

The only change to ensure physical consistency is to alter the lower bound of $\tilde{M}_z$ ($\tilde{M}_z = M$ otherwise) such that the z update obeys $z_{i+K} \geq \mathrm{SWE}_{i+K}$ to enforce $z \geq \mathrm{SWE}$, so SWE_{i+K} was calculated first before dz/dt. This permits comparison with other models without inputting observational SWE.

To compare the neural model to established parameterizations, the Snow17 temperature-index model (Anderson 2006) was implemented and evaluated on the same data. Snow17 (designed for modeling runoff) models SWE and infers depth through Anderson's density parameterization, and it can also assimilate observed SWE, allowing comprehensive comparisons. We performed two comparisons: 1) M (the one-network depth parameterization) against Snow17 with both assimilating observational SWE and 2) the two-network standalone model $\tilde{M}$ (subcomponents $\tilde{M}_z$ and $\tilde{M}_{\mathrm{SWE}}$) against Snow17, both using only meteorological inputs. To avoid confusion, Snow17 predicting versus assimilating SWE is labeled as "SN17" and "SN17O," respectively. We used the Snow17 parameters from the Wang et al. (2022) "SN17-B-CONUS" model, calibrated over CONUS to account for regional climates. Both models were restarted with accurate depth when resets occurred, and computational benchmarking is compared in Table C4. Significance in differences between RMSE metrics was evaluated via a Wilcoxon signed-rank test (Wilcoxon 1945) with $p = 0.05$, which does not assume normal distributions or variance homogeneity.

In addition to depth, bulk density time series can be generated from z and SWE data and model outputs according to Eq. (1) and compared. The data are discrete while the model outputs are continuous, so densities were compared at sites with collocated SWE and z sensors (see appendix B) only when both models and data yielded physical values ($0 < \rho_{\mathrm{snow}}/\rho_{\mathrm{water}} < 1$). Counts of "false nonsnowpacks" (models show $z = 0$ while $z > 0$ in the data) and "false snowpacks" (models show $z > 0$ while $z = 0$ in the data) were recorded, along with instances of unphysical densities.

To further assess physical consistency, the estimated rates $d\mathrm{SWE}/dt$ by $\tilde{M}_{\mathrm{SWE}}$ were compared to the snowfall rate data P_{snow}. When the air temperature is below freezing ($T_{\mathrm{air}} < 0$), the conservation of mass implies these rates should be roughly equivalent, limited by factors such as runoff, sublimation, snow transport, and data precision. While it is possible to directly impose conservation of mass as a threshold in this framework, forgoing this specific bound allows for an investigation of the model's ability to represent physical constraints beyond those explicitly encoded.

Assessing physical consistency in the model directly per feature is nuanced due to strong correlations among the

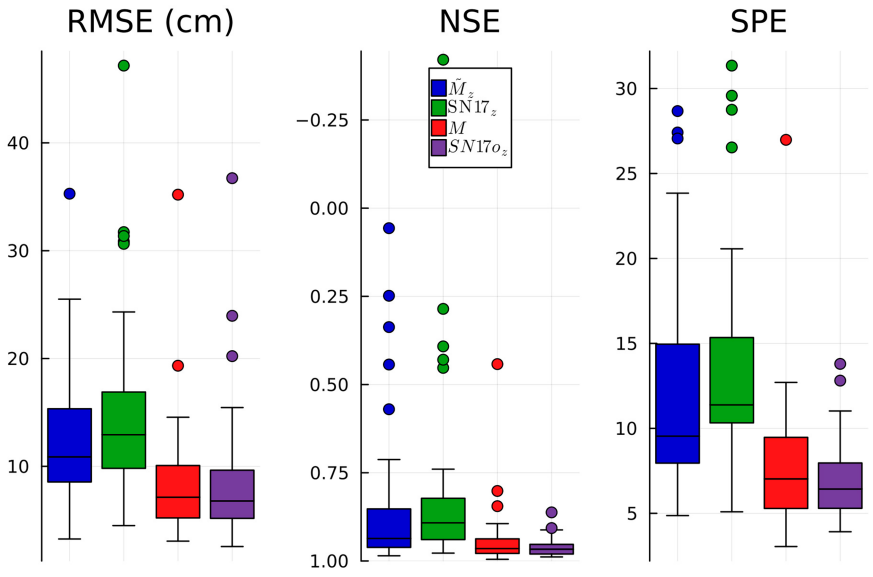

FIG. 4. Performance of M and $\tilde{M}$ against Snow17 with (SN17O) and without (SN17) observational SWE data for generating z time series over the 44 validation sites. RMSE indicates root-mean-square error, NSE indicates the Nash–Sutcliffe efficiency, and SPE indicates the average L1 error normalized to the average nonzero depth, to measure percent error. Boxes outline the 25%–75% quantiles, with the bar at the median, while whiskers mark the extremes and dots indicate outliers, which lie beyond 1.5 times the IQR (box width) from the box. Vertical axis limits are chosen to show all data points.

inputs, which confound most interpretation methods like partial dependence, permutation, local interpretable model-agnostic explanations (LIME), and tractable Shapley additive explanations (SHAP) values (Molnar 2022). To isolate feature effects on model output, we calculated first-order accumulated local effect (ALE) plots (Apley and Zhu 2020). This method shows the average change in model output accumulated over sequential bins of feature values. For feature X at grid point x_i, data in the window $[x_i - \Delta, x_i + \Delta]$ are evaluated setting $X = x_i - \Delta$ and $X = x_i + \Delta$, storing the average of the differences $\Delta M_{x_i} = \overline{M(x_i + \Delta) - M(x_i - \Delta)}$. The final (centered) ALE value $\overline{\Delta M}$ at x_i is the sum over all ΔM_{x_k} with $x_k \leq x_i$, minus the average of all uncentered ALE values. This isolates changes solely from feature variations, and bins are defined by quantiles to ensure equal data instances in each window. The shape and slope of the ALE curve are more pertinent for interpreting physicality than the offset, especially when the feature distribution is skewed. The range of $\overline{\Delta M}$ over the feature indirectly measures feature importance in influencing model output, as the ALE value can be interpreted as a departure from the average model prediction. Further information for interpreting ALE plots can be found in Molnar (2022).

3. Results

a. Depth time series

Over validation sites, neural configurations performed similarly to Snow17 independent of including observed SWE (Fig. 4). Both models exhibit similarly tight spreads, though M exhibited larger spreads than SN17O when utilizing SWE data. Conversely, when simulating SWE, the spread was larger for SN17 compared to the coupled neural model $\tilde{M}$, which is more indicative of usage under SWE uncertainty or provision of SWE within a separate model. Statistically significant improvements in RMSE were found for both SWE and z modeling by $\tilde{M}$ compared to SN17 ($p = 0.029$ and $p = 0.0006$, respectively). No significant difference was observed for M against SN17O using SWE data for depth parameterization alone ($p = 0.673$), which is unsurprising as both models were calibrated for performance over the locations represented in these data.

The performance of the models over all testing sites is summarized in Fig. 5, with example time series in Fig. 6 and medians and averages reported in appendix C (Table C2). Although the performance distributions are nonsymmetric, both mean and median metrics are important for gauging potential (median) and consistency (mean) for generalizing to out-of-sample regimes. The neural configurations showed tighter performance spreads across testing sites and climates not included in the training data, despite no recalibration. Average neural NSE was 0.87 and 0.94 (without and with SWE data, respectively), compared to 0.35 and 0.78 for Snow17—median differences were smaller, but still favored neural models. Snow17's mean SPE was 15% even with observational SWE, comparable to other established models without site-specific calibration (Vionnet et al. 2012; Brun et al. 2013; Viallon-Galinier et al. 2020; Luijting et al. 2018; Ebner et al. 2021; Meloche et al. 2022; Gao et al. 2021; De Michele et al. 2013), and nearly

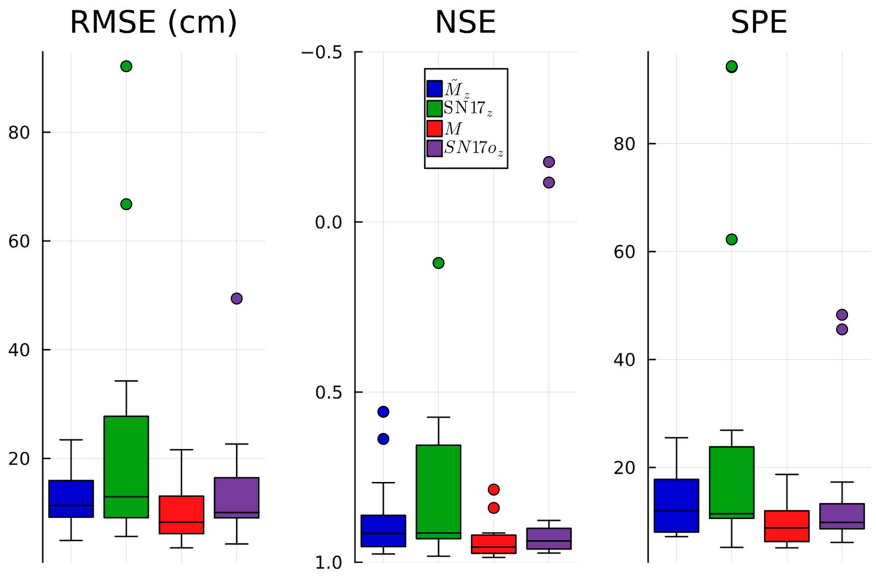

FIG. 5. Performance of the neural parameterization against Snow17 over the 14 testing sites in depth time-series generation. Labeling convention follows that in Fig. 4. One outlier for SN17 with an NSE of −2.8 is not shown on the plot to aid in the scaling of the other values.

doubled to 28% without SWE data. In contrast, the neural parameterization maintained mean SPE under 15%, increasing modestly from 9.5% to 13.4% without SWE data, demonstrating greater robustness and consistency out-of-sample despite the simplicity of its predictive component and lack of past snow depth state storage.

The parameterization M demonstrates strong generalizability, with similar performance and tighter spreads across test and validation sites, without retuning. Although the neural model ($\tilde{M}$) did not exhibit significant RMSE improvements over Snow17 on out-of-sample data at the $p = 0.05$ level ($p = 0.14$ for z and $p = 0.27$ for SWE), it did outperform Snow17 as a parameterization ($p = 0.01$). This suggests that when paired with a reliable SWE predictor within a larger model, the neural approach offers advantages over prevailing alternatives in extrapolating to uncalibrated locations.

b. Density time series

For validation and testing sites with collocated z and SWE measurements (see appendix B), Table 1 compares time-series statistics for derived bulk density, and Fig. 7 displays estimates from the same data as Fig. 6. RMSE values reflect relative error since density is normalized by ρ_{water}. Both SN17 and $\tilde{M}$ lack SWE discretization, inflating errors for small snowpacks compared to their assimilated counterparts. Neural unphysical densities source from a lack of constraint enforcing SWE $= 0 \implies z = 0$ (in both $\tilde{M}, M$) or $z >$ SWE (in M), though different constraint choices could eliminate this. Snow17 is roughly 3 times better at mitigating false snowpacks with lower RMSE (less extreme errors), but our models are roughly 3 times better at reducing false snow absence and exhibit improved individual density predictions, crucial for estimating other snowpack properties. Thus, the advantage between models depends on which features are prioritized. Otherwise, the two perform similarly, despite M being designed for snow depth while the comparable parameterization within Snow17 is explicitly formulated for density. Among evaluated sites, the best and worst performances by M were better than those of SN17O.

c. Predicted dz/dt

Figure 8 shows a histogram comparing predicted versus true dz/dt values from M (Pearson correlation $r = 0.78$) and $\tilde{M}$ ($r = 0.77$) during time-series generation. Both still display a tendency to underpredict extreme values. This is likely from exposure to much more training data with small dz/dt, which could result in better predictions of small values at the expense of extremes without applying other methods like class balancing.

d. Physical behavior of model

For SNOTEL instances with $T_{\text{air},i} < 0$, $\tilde{M}_{\text{SWE}}$ violated the $d\text{SWE}/dt \leq P_{\text{snow}}$ condition only 0.6% of the time, with all violations under 1 mm day^{-1} and 95% under 0.75 mm day^{-1}. About half the instances showed $d\text{SWE}/dt < P_{\text{snow}}$ (prescribed bounds yielded equivalence in the rest), but the mean residual of these was −2 mm day^{-1}, with 95% above −5 mm day^{-1}, aligning with typical sublimation rates (Spehlmann et al. 2025; Liu et al. 2024) and effects not explicitly modeled. Given the SNOTEL data precision (2.54 mm) and negative average temperatures obscuring instantaneous positive temperatures, these results suggest strong adherence to mass conservation, even without explicitly enforcing it (as simple as choosing $f_+ = P_{\text{snow}}$ from section 2b for $\tilde{M}_{\text{SWE}}$), highlighting the model's ability to learn physical representations from data beyond those prescribed.

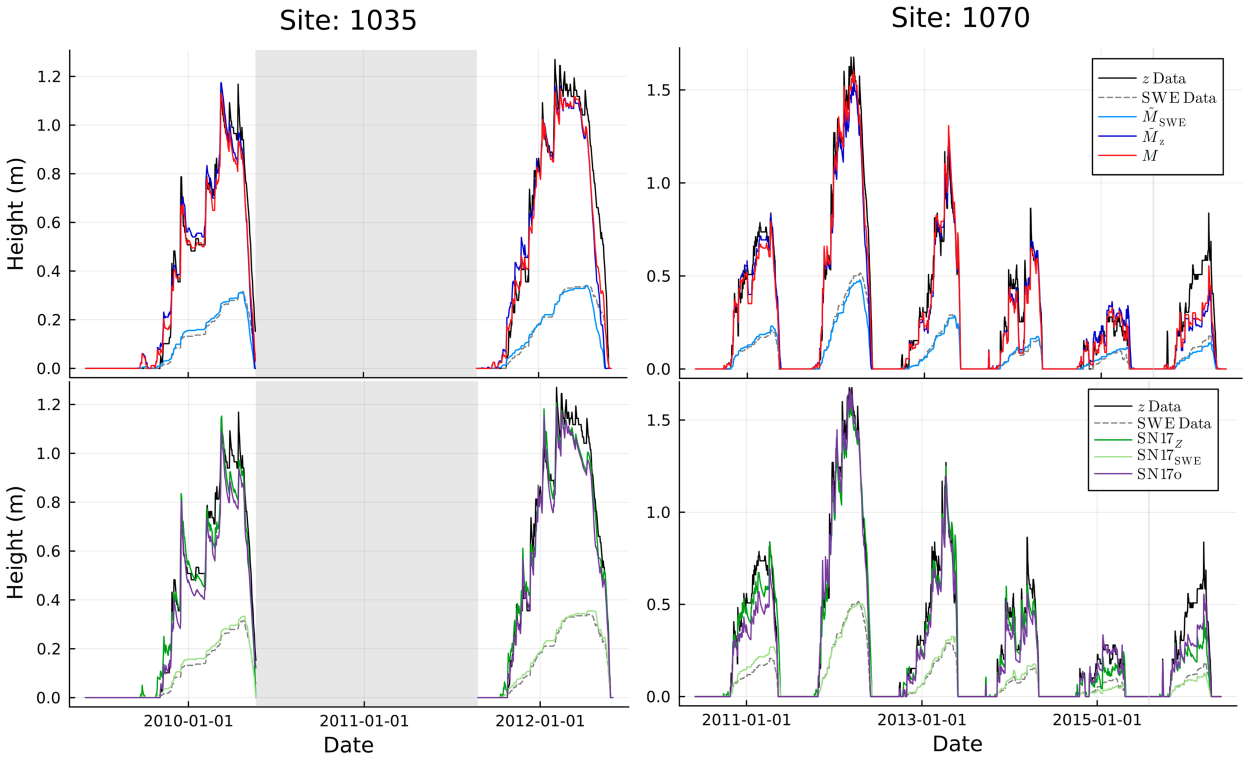

FIG. 6. Simulated z and SWE time series over two of the Alaskan testing sites. Data gaps (missing or cleaned data) are shaded in gray. All models perform similarly, though the neural models seem to perform poorly whenever Snow17 also performs poorly, like in the winter of 2016 at site 1070. This suggests the neural parameterization is at least an efficient surrogate for Snow17 or could insinuate a data inconsistency if both models fail similarly.

The ALE plots for each feature are shown in Fig. 9. The average prediction of M is ≈0, with negligible centering offsets, indicating a tendency to predict negative dz/dt after $T_{air} > 0$. The parameterization M also exhibits a linear relationship in dz/dt with P_{snow} and stabilization at low T_{air} (where a minimum snowfall density would emerge), all aligning with physical expectations. Additionally, decreases in dz/dt are observed with increased solar radiation and wind speeds, reflecting the model's learning of understood destructive processes. However, data availability and artifacts may influence these results; for instance, while higher relative humidity could lead to more condensation and surface melt, it also lowers the estimated snowfall fraction under the data preparation used. Moreover, snow accumulation on sensors during heavy snowfall may produce saturated humidity readings linked to large dz/dt, and location biases may affect the value ranges.

The parameterization M can be directly queried over "slices" of its multidimensional input space, facilitating the exploration of the model's behavior in anticipated input regions that may not be observed in the current data. Unlike partial dependence plots, these slices are not averaged over training data ranges, allowing them to display outputs in regimes of inputs that are not physically viable. Additionally, patterns observed in one "slice" may not be conserved across others. Figure 10 illustrates two such slices, with areas lacking data slightly masked.

TABLE 1. Results of bulk density time-series generation for each model. The median score over all validation and viable testing sites is presented, and scores are derived from all predictions of physical densities during observed snowpacks, or otherwise tallied in the presented counts. The integer count gives the median number of occurrences across sites, while the percentage normalizes each count against the length of the time series.

Parameter	$\tilde{M}$	M	SN17	SN17O
RMSE (%)	9.1	9.5	7.4	7.6
Bias (%)	−1.9	−1.8	0.6	0.9
MPE (%)	11.3	12.9	14.1	14.9
False nonsnowpacks	4 (0.13%)	2 (0.08%)	14 (0.68%)	0 (0%)
False snowpacks	183 (8.6%)	136 (6.1%)	45 (3.2%)	2 (0.10%)
Unphysical points	3 (0.17%)	3 (0.12%)	0 (0%)	0 (0%)

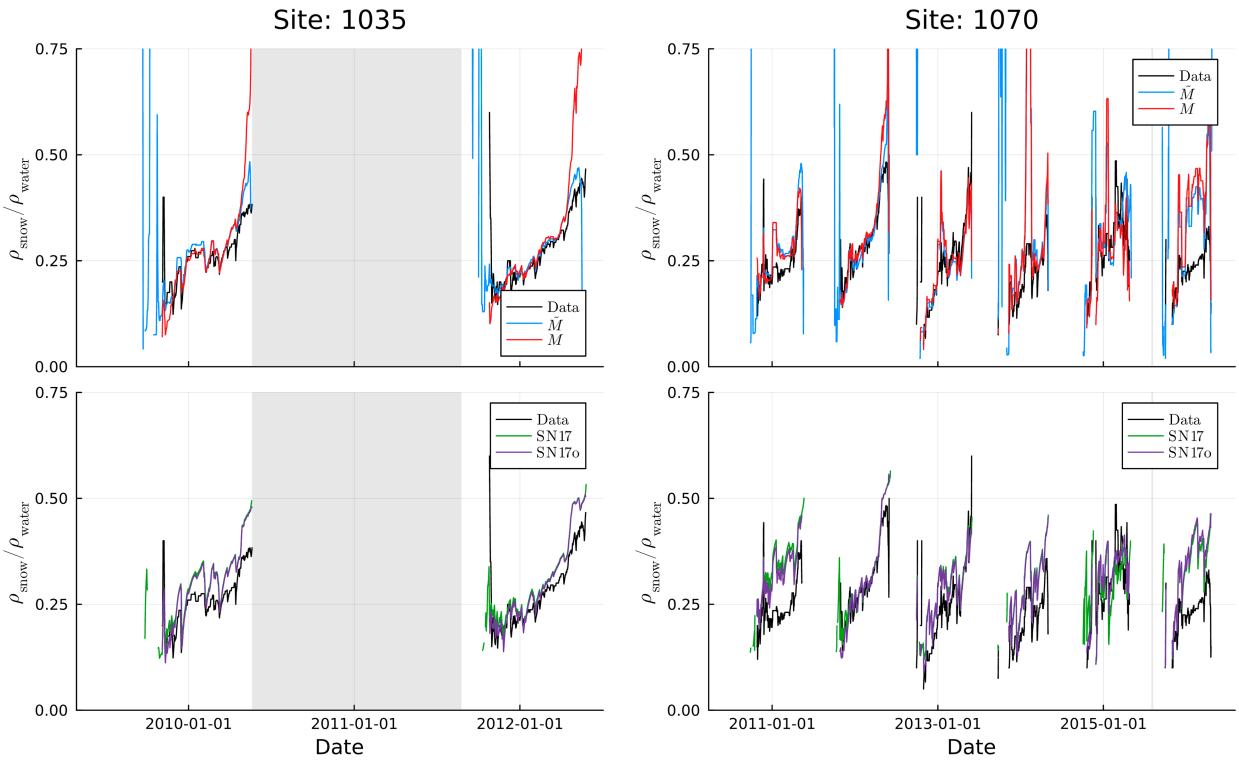

FIG. 7. Bulk density time series (normalized against ρ_{water}) over the same series as Fig. 6. Poor representation at the start/end of the season for small snowpacks is exacerbated by data discretization and lack thereof in the models, particularly in $\tilde{M}$ and SN17, or from simpler constraints in M and $\tilde{M}$. Such fluctuations severely skew the NSE and SPE metrics and confound their measure of model performance. Time series are only shown where models and data show a relative density between 0 and 1.

In Fig. 10a, contours reflect physical expectations, with even spacing indicating dz/dt is linear in P_{snow}, T_{air} becoming the dominant feature around $T_{air} = 0$, and the constraint $\Delta z \geq -z$. Figure 10b examines how T_{air} and insolation impact snow depth at zero snowfall, showing depletion begins once $T_{air} > 0$ and with increasing R. All output values at zero snowfall remain nonpositive, adhering to the prescribed threshold. This slice demonstrates limited sensitivity to solar insolation at low R, suggesting T_{air} is the dominant variable. This insensitivity may arise from high snow reflectivity at low incidence angles (and normal sensor orientation), shading effects reducing melting until higher radiation levels are exposed, or latent melting effects prevailing at low irradiance. Positive feedback loops where accumulating surface melt alters albedo could explain the transition in this slice. Overall, the model aligns with expected physical behavior over available data, though incorporating more data into extrapolated (shaded) regions could enhance universality.

e. *Generalizability*

To assess generalizability, Fig. 11 shows elevation versus mean nonzero snow depth $\overline{z}_+$ for all training and testing sites, similar to Fig. 3. Sites are colored by the performance of M for SPE, RMSE, and NSE on z time series, along with density RMSE. The model succeeds comprehensively, with SPE errors under 20% for nearly all sites and most under 10%, while density errors remain predominantly below 15% RMSE. It performs comparably on testing data from different elevations and climates, indicating robust generalizability, even for density calculations. No discernible trend with elevation appears, corroborating generalizability rather than elevation-induced effects.

f. *Finer-resolution predictions*

Time units only appear as rates in M via precipitation, the lower bound, and the output. By predicting rates dz/dt instead of accumulated dz, M can be evaluated at varying time steps without retraining, by merely resetting the constraint function's scaling of $1/\Delta t$ without altering the trained predictive weights (which were trained only once in section 3a, with $\Delta t = 1$ day). This flexibility allows testing at resolutions up to data limits, which for Kühtai is a maximum resolution of 15-min intervals. While subhourly time resolutions are rarer in larger Earth models, they are used in land models and will become increasingly frequent as research advances in fine-scale land–atmosphere interactions (Schär et al. 2020; Ban et al. 2021). Similarly, while multiday resolutions are rare, data availability can necessitate their use. Benchmarking across scales above and below the daily resolution used for training offers insights into the model's limits and potential applications. However, beyond a week, average input variables do not effectively capture critical input dynamics (e.g., monthly average T_{air}

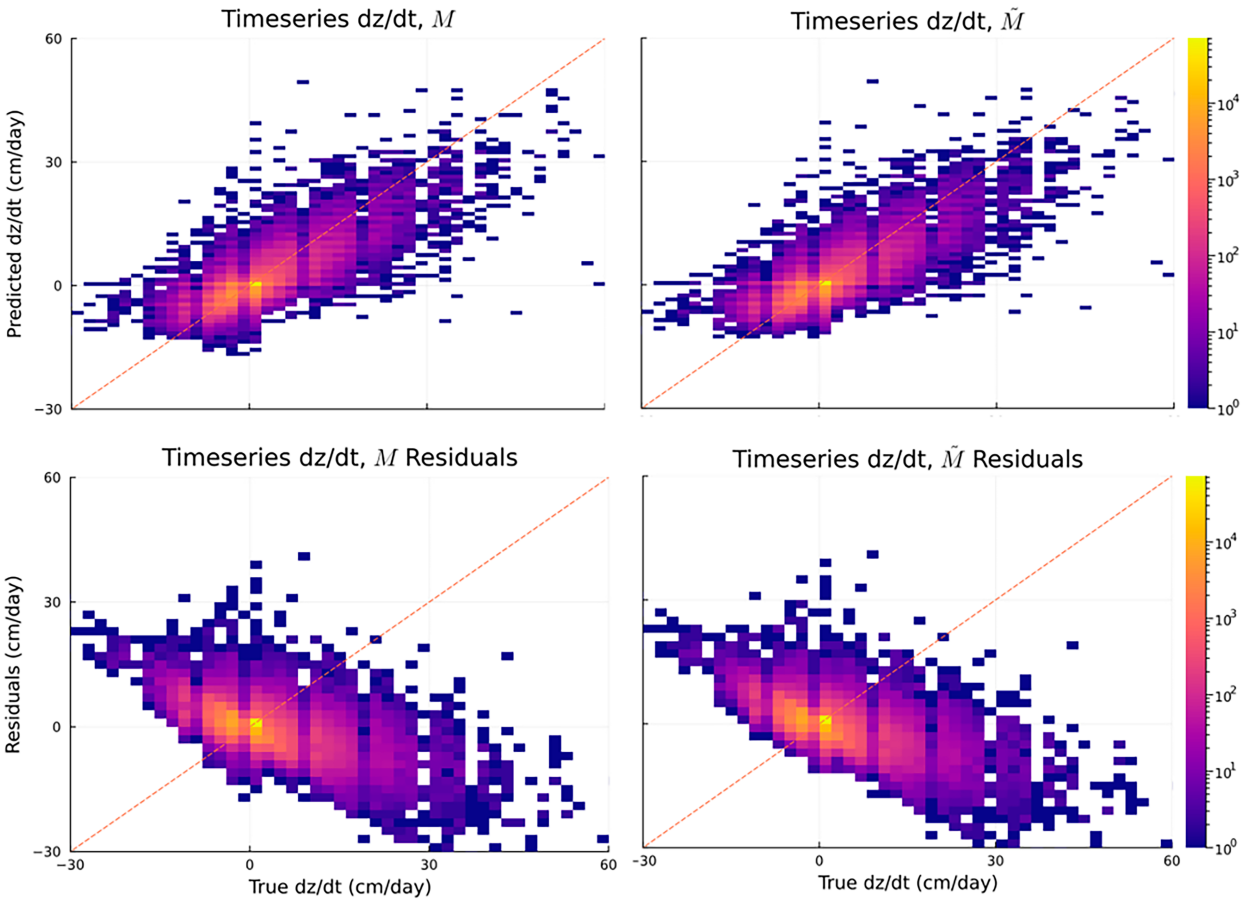

FIG. 8. Predicted vs true dz/dt and residuals against the modeled target by the M and $\tilde{M}$ models. Both models continue to underpredict extremes, despite the bias given to extremes during training. The magnitude of data in small dz/dt ranges relative to extremes could contribute to this phenomenon.

would not reliably indicate time above freezing), leading to suppressed variations that undermine meaningful outputs, and alternative model choices would be more appropriate. Results of repeating the time-series generation with M as described in section 3f for the Kühtai data with different values of Δt in the constraint but identical weights otherwise, spanning 15-min to weekly resolutions, are shown in Fig. 12 and Table 2.

The time series shown at this site all exhibit a low bias; however, for all resolutions, M achieved over a 40% reduction in RMSE compared to Snow17 (not tabulated). RMSE increases at time steps outside the daily interval used for training, though all subhourly resolutions yield nearly identical results without further trends. This performance loss may arise from the extreme values of dz/dt and precipitation observed at higher resolutions. While daily data show gradual snowpack increases, finer resolutions might capture the same deposition over a few hours, leading to dz/dt values 10–20 times larger than those in the training data. Conversely, weekly averages smooth out extreme events, which can reduce the variance of outputs. Overall, M demonstrates an ability to transfer across temporal resolutions, though performance would likely improve with a broader range of dz/dt training values (for instance, incorporating both hourly and daily data).

4. Discussion

This study explored a simple, versatile data-driven framework to enhance physical parameterizations, focusing on generalizable snow parameterizations for climate modeling applications. Many choices were results driven and informed by data availability, such as selecting widely measured variables to increase model applicability, though other choices could likely enable the representation of additional processes. Data requirements limited sources to the SNOTEL network, which, while useful for local relationships, has quality issues compared to validation sites such as Col de Porte and Kühtai (Meyer et al. 2012, 2023), fails to represent large-scale heterogeneity effects, especially in mountain regimes (Meyer et al. 2023), and lacks coverage of extreme conditions in scarcely sampled tundra and taiga biomes. The approach's ability to extend to these unsampled terrains or perform on coarser grids with explicit large-scale effects within Earth system model remains

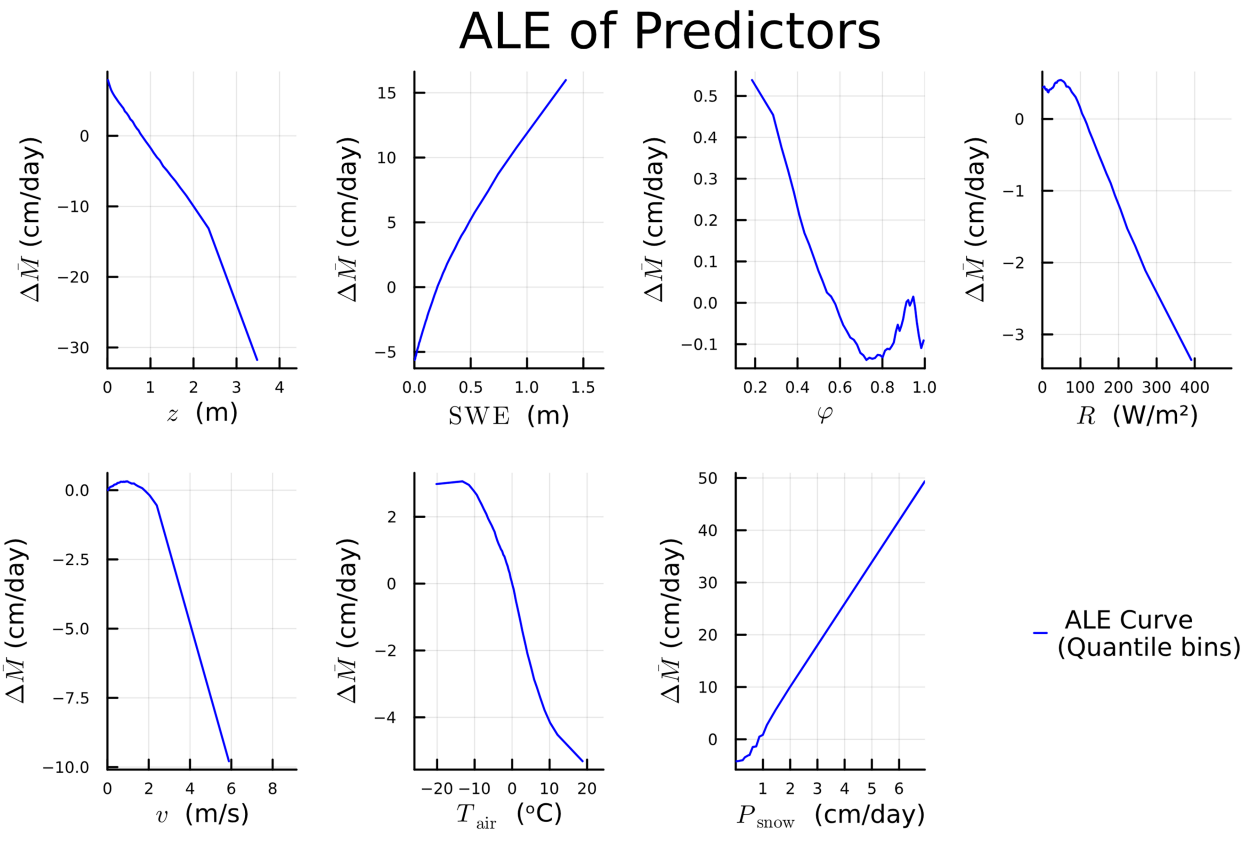

FIG. 9. ALE plot for M's predictive features. The ALE mitigates the effects of feature correlations, providing a rough indication of feature importance in determining M's output magnitude and a visual measure of physicality. The horizontal axes mark the range of each feature, while the vertical axes show the change in M relative to the average prediction. Curves are binned by quantiles so each bin has at least 50 samples.

untested and will be the subject of a future paper. Further adaptation of these models in a world of growing data volume, frequency, and quality offers an exciting opportunity for future research.

The neural parameterization M generalizes well across locations, a crucial feature for both global consistency and local-scale modeling, particularly as climate change skews site statistics to become "new" locations. Snow17 struggles with such shifts,

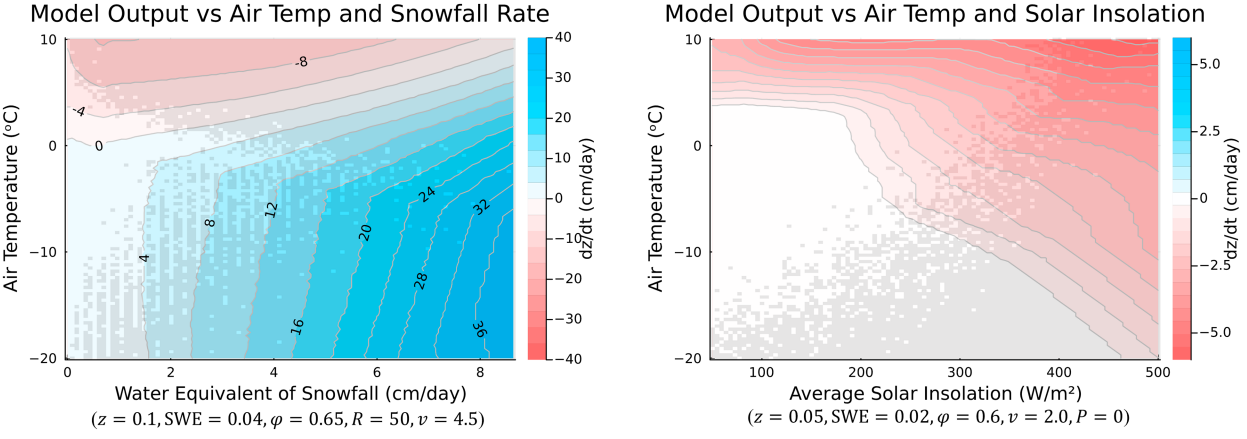

FIG. 10. Outputs of M for two snow states, with other inputs held constant. In each case, one threshold ($\Delta z > -z$ or $dz/dt < 0$ for $P_{\text{snow}} = 0$) is visible. "Shaded" or "masked" pixels (partially greyed out) indicate areas of the visible parameter space with ranges not represented in the data, from being either unphysical or unobserved.

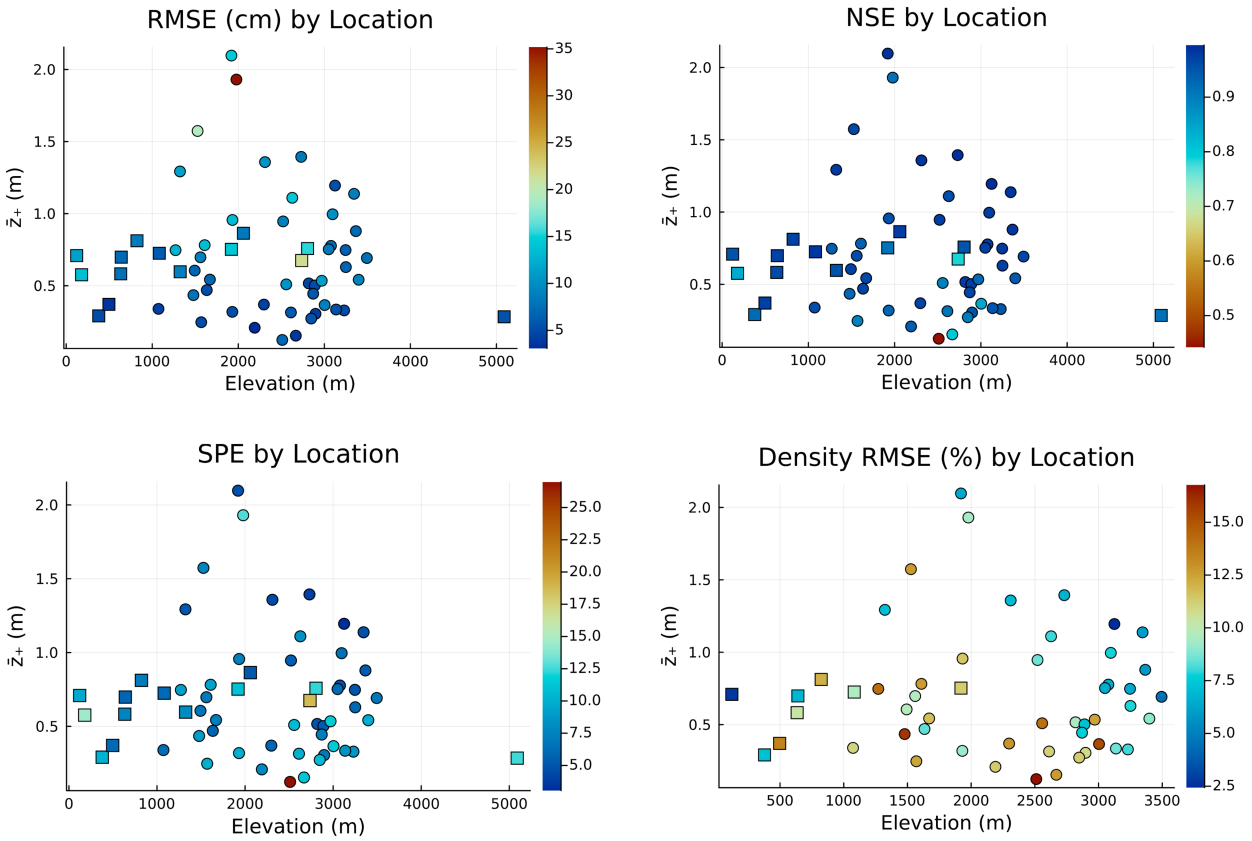

FIG. 11. Performance of M across all training and testing sites with regard to RMSE, NSE, and SPE of depth time series, as well as RMSE for density time series over the sites where density was evaluated (see appendix B). Testing sites are marked with squares instead of circles.

limiting its application on longer time scales (Meyer et al. 2023; Boone and Etchevers 2001). Our physics/ML hybrid approach offers a viable alternative. Generalizability is key for ML in climate models. For instance, Yang et al.'s (2020) random forest model applied to Chinese sites showed high out-of-sample biases and SPE compared to our model, despite similar RMSE. Their ML approach and that of others like Duan et al. (2024) or Yang et al. (2022) rely on location-specific variables (slope, aspect, topographic/vegetative indices, etc.) or features like historical averages and microwave measurements (Tanniru and Ramsankaran 2023; Song et al. 2024; Cui et al. 2023; Vafakhah et al. 2022; Yang et al. 2022), which hinder their usage within Earth system models. In contrast, Wang et al. (2022) simulated SWE with recurrent neural networks using physical inputs with similarly high NSE scores and moderate generalizability but required over 240 previous states for updates. Duan et al. (2024) also used 180 days of forcing data per SWE prediction for various models at SNOTEL sites, requiring training times of 5–26 h and hours of simulation time on a GPU. Our memoryless neural ODE model achieves comparable results to Wang et al. (2022) with guaranteed consistency with physical bounds. It exhibits improved median MAE, RMSE, and NSE against all models from Duan et al. (2024), with fewer inputs and significantly fewer computational resources (see Table C4). Steele et al. (2024) used the same inputs as ours for a standalone ML model and a postprocessing model for physical models. Both yielded higher SWE RMSEs (6 and 13 cm versus ~4 cm in our model; see Table C2) and poorer generalization, though they produced a slightly better derived density RMSE (implying z errors were similarly scaled to SWE errors). Both models required the addition of a binary snow-presence variable to reset unphysical summer snowpacks, while our approach naturally adheres to prescribed bounds, eliminating such unphysical departures.

The best hyperparameters for time-series generation differed from those for direct regression, underscoring that better individual dz/dt predictions do not guarantee better accumulated seasonal time series. This further supports that our choices are going beyond merely matching magnitudes and instead summarizing universal, memoryless physical processes. Notably, the parameterization faltered only when new climates introduced target magnitudes absent in the training data, rather than when locations presented different input feature magnitudes. While output magnitude extrapolation is limited as with many data-driven models, the input generalizability is a less common result, highlighting the benefits of enabling physical consistency. This underscores the need for more widespread snow sensing, particularly in extreme climates, to improve the predictive power of such models.

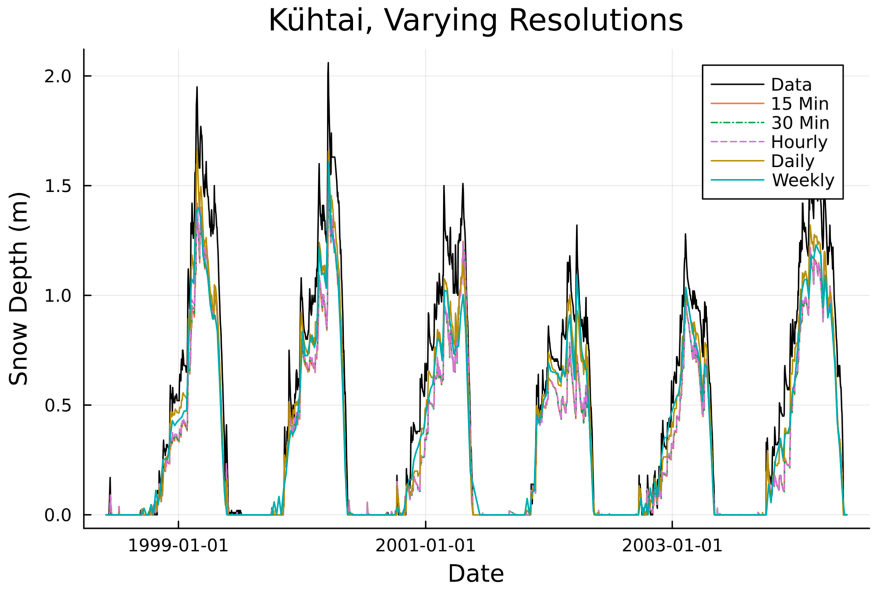

FIG. 12. Output of M at different resolutions for a subset of the site data from Kühtai, Austria. The graph overlays outputs at different resolutions for direct comparison. Sub-hourly (15, 30 min) curves are nearly indistinguishable from the hourly curve.

The framework's application of prescribed constraints around black-box models like neural networks enables easy modification of constraints or input variables. This interchangeability supported rapid experimentation and prototyping and is synergistic for integration as a "plug-and-play" model that can adapt to available inputs and operational constraints. The approach provides linear scaling in input size with low computational overhead, which can reduce computational budgets while maintaining or improving accuracy compared to advanced models, permitting simulations longer into the future or over finer grids. It can demonstrably act as a standalone predictive tool wherever inputs can be measured or inferred, which could be through observations, remote sensing, weather forecasts, or coupled models. This framework could be used, for example, for forecasting applications such as weekly skiing or hiking terrain predictions from weather forecasts, or tested for water supply simulation given snowpack data. The model structure is a strong candidate for simulating many types of constrained physical systems beyond snow, offering further avenues for investigation (see appendix A).

Beyond testing the neural parameterization M globally in a coupled setting within an entire climate model, other future directions of research could involve adapting M to continuous neural ODE structures (Chen et al. 2018) or more general time-stepping schemes. While this study focused on localized relationships and comparisons with similarly formulated standards, it did not address large-scale heterogeneity effects. Future adaptations to directly incorporate these effects might aid usage within coarser-scale (~100 km) models. Additional improvements could involve integrating data from the National Operational Hydrologic Remote Sensing Center (NOHRSC) or Meteorological Assimilation Data Ingest System (MADIS), or detailed snow layer data to simulate temperature profiles. Alternative training strategies, such as using time-series error as the loss function or gradient-free update rules to bypass recursive time-series gradient issues, could also be explored.

TABLE 2. Performance of M at varying resolutions for z time-series generation. Error jumps beyond the daily training resolution, but performance remains near constant between hourly and 15-min resolution.

Statistic/resolution	Weekly	Daily	Hourly	30 min	15 min
RMSE (m)	0.1762	0.1437	0.2008	0.1989	0.1999
NSE	0.759	0.806	0.708	0.710	0.709
SPE (%)	23.90	19.54	30.9	31.0	31.1

5. Conclusions

Using a location-agnostic and physically constrained neural ODE framework to parameterize the rate of change in snow depth, we were able to simulate seasonal snow depth with a median error of 8.8% across sites with varying climates and elevations, including some not seen during training. Though the parameterization was trained with daily data, it shows an ability to perform with moderate accuracy at other temporal resolutions without additional retraining of the model; however, retraining with higher-resolution training data may lead to further improvement. The parameterization's structure reduces computational overhead while maintaining performance in depth simulation at the level of established, cutting-edge, or more detailed models. The design is conducive for usage in prognostic models or can be adapted to alternatively predict variables such as SWE. When driven solely by meteorological data as a standalone model, the parameterization framework can recreate seasonal time series with comparable error without

retraining or site calibration—an improvement over other established models. In most cases, it matches or outperforms the Anderson Snow17 model in simulating seasonal snow depth, offering an efficient formulation for use within physical models and an alternative to prevailing parameterizations.

The proposed framework demonstrates the potential for a wide array of applications for both long-term climate simulations and short-term forecasting applications. The means of enforcing hard constraints structurally provides a simple but powerful technique for predictive modeling that can be applied beyond snowpack modeling to different climate processes or physical parameterizations.

Acknowledgments. We thank Marie Dumont for insightful discourse on process-based snow models, the SNOTEL effort, and the Kühtai, Col de Porte, Reynolds Mountain East, Sodankyla, Rofental, and Yala Basecamp teams for their data. A. C. was supported by the AI4Science initiative at the California Institute for Technology and a Department of Defense National Defense Science and Engineering Graduate (NDSEG) Fellowship. This work was generously supported by Schmidt Sciences, L.L.C., and the Resnick Sustainability Institute. The authors thank Jeffrey Coyle, Jaz Ammon, Joseph Kral, Matt Warbritton, and Daniel Tappa for help in verifying SNOTEL sensor placement and Yuan-Heng Wang for sharing calibrated Snow17 parameters.

Data availability statement. The SNOTEL data utilized in this study were available via the National Water and Climate Center, which lies under the United States Department of Agriculture. Data reports of the SNOTEL data were generated using the online portal found at https://www.nrcs.usda.gov/wps/portal/wcc/home/. The data from Col de Porte (Lejeune et al. 2019) can be found at the Observatoire des Sciences de l'Univers de Grenoble DOI portal at https://doi.osug.fr/public/CRYOBSCLIM_CDP/CRYOBSCLIM.CDP.2018.html, and data from Kühtai (Krajči et al. 2017) can be found as the supplemental material from https://doi.org/10.1002/2017WR020445. Raw data from Sodankyla (Essery et al. 2016) can be found from the Intensive Observation (sensors 8, 11) data portal at https://litdb.fmi.fi/ioa.php and automated weather station (sensor 15, portal at https://litdb.fmi.fi/luo0015_data.php). Reynolds Mountain East (Reba et al. 2011) data were obtained from an ESM-SnowMIP repository https://www.geos.ed.ac.uk/~ressery/ESM-SnowMIP.html. Raw Yala Basecamp (Stigter et al. 2021; Shea et al. 2015) data were retrieved from https://rds.icimod.org/Home/DataDetail?metadataId=26859 and https://rds.icimod.org/Home/DataDetail?metadataId=1972554. Rofental (Warscher et al. 2024) data were retrieved from https://datapub.gfz-potsdam.de/download/10.5880.FIDGEO.2023.037-MNveB/. The data in this study were processed from these sources. Code for scraping and cleaning SNOTEL data and tutorials for data retrieval and training/modifying the neural models are available at https://clima.github.io/ClimaLand.jl/dev/generated/standalone/Snow/base_tutorial/. CSVs of the training/testing data are also available here, as a quality-controlled fully observational dataset of physical variables for calibration and ML applications.

APPENDIX A

Threshold Constraint Layers

a. Defining threshold constraint layers

Since $\mathrm{ReLU}(x) = \max(x, 0)$, we can reexpress the minimum and maximum functions as

$$\max(x, y) = y + \mathrm{ReLU}(x - y) = \mathrm{ReLU}(y) - \mathrm{ReLU}(-y) + \mathrm{ReLU}(x - y) = \max(y, x), \quad \text{(A1)}$$

$$\min(x, y) = y - \mathrm{ReLU}(y - x) = \mathrm{ReLU}(y) - \mathrm{ReLU}(-y) - \mathrm{ReLU}(y - x) = \min(y, x). \quad \text{(A2)}$$

Then, for a model output p and any construction f serving to threshold p, the bounds $\max(f, p)$ or $\min(f, p)$ can be explicitly implemented with a single depth-3 fixed-weight layer with no biases acting on input $[p, f]^{\mathrm{T}}$ with ReLU activation, followed by an accumulation with no activation:

$$\begin{bmatrix} \pm 1 & 1 & -1 \end{bmatrix} \times \mathrm{ReLU}\left(\begin{bmatrix} \pm 1 & \mp 1 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \times \begin{bmatrix} p \\ f \end{bmatrix} \right) = \mathbf{A}_{1\pm}^{\top} \mathrm{ReLU}\left(\mathbf{A}_{2\pm}^{\top} \begin{bmatrix} p \\ f \end{bmatrix} \right), \quad \text{(A3)}$$

where $+$ indicates $\max(f, p)$ and $-$ indicates $\min(f, p)$, and the ReLU acts elementwise. Equation (A3) offers a simple formulation, but the symmetry of the maximum and minimum functions permits resonant structures—different weights yielding the same output. If the signs of p or f are unknown, both $+f$ and $-f$ (or $+p$ and $-p$) must be passed through the ReLU, along with $p - f$, to preserve all necessary information. However, if f is always nonnegative (or nonpositive), the layer depth can be simplified from three to two, as passing $-f$ (or $+f$) becomes redundant and its ReLU always evaluates to zero. Similarly, if the threshold obeys $f \geq C$ (or $p \geq C$) for some constant C, a similar reduction is possible by including a bias term along with $A_{2\pm}$ and $A_{1\pm}$. The same reductions apply if p also exhibits similar properties. Figure A1a illustrates the generalized structure for one-sided threshold constraints on a predictive component discussed in section 2b.

Likewise, for a simultaneous upper bound f_+ and lower bound f_- on p for any constructions f_+, f_- satisfying $f_+ \geq f_-$, we have

$$\max[\min(p, f_+), f_-] = \mathrm{ReLU}(f_-) - \mathrm{ReLU}(-f_-) + \mathrm{ReLU}(\alpha), \quad \text{(A4)}$$

where

$$\alpha = \mathrm{ReLU}(f_+) - \mathrm{ReLU}(-f_+) - \mathrm{ReLU}(f_-) + \mathrm{ReLU}(-f_-) - \mathrm{ReLU}(f_+ - p), \quad \text{(A5)}$$

so the threshold can be explicitly implemented with a sequence of two fixed-weight layers containing no biases acting

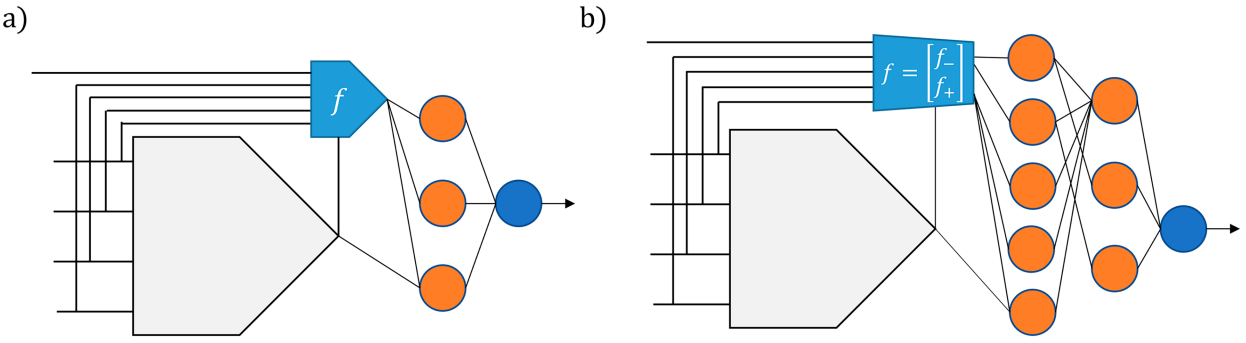

FIG. A1. In all graphics, inputs are on the left, with predictions on the right, and the gray pentagon depicts the predictive component from Fig. 1. Knowledge of function f can simplify layers by removing certain orange (ReLU) nodes, with resonant structures possible from the symmetry of the maximum and minimum functions. Black weights are fixed at $+1$ or -1, with no biases or training. (a) General structure for a one-sided constraint (maximum or minimum) on the prediction. (b) General structure for a two-sided constraint (enforced range) on the prediction.

on input $[p, f_+, f_-]^{\mathrm{T}}$, followed by an accumulation with no activation:

$$\begin{bmatrix} 1 & 1 & -1 \end{bmatrix} \times \mathrm{ReLU}\left[\begin{bmatrix} -1 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \mathrm{ReLU}\left(\begin{bmatrix} -1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{bmatrix} \times \begin{bmatrix} p \\ f_+ \\ f_- \end{bmatrix}\right)\right], \tag{A6}$$

$$= \mathbf{A}_{1+}^{\top}\mathrm{ReLU}\left[\mathbf{A}_3^{\top}\mathrm{ReLU}\left(\mathbf{A}_4^{\top}\begin{bmatrix} p \\ f_+ \\ f_- \end{bmatrix}\right)\right], \tag{A7}$$

which takes advantage of the identity $\mathrm{ReLU}[\mathrm{ReLU}(x)] = \mathrm{ReLU}(x)$. Like the one-sided threshold example, many resonant structures exist according to symmetry, and bounds on the thresholds or p permit layer reductions. Figure A1b shows the generalized structure for a two-sided threshold constraint function f outputting f_+, f_- on the predictive component given in section 2b.

These constraint structures can adapt to any functional constraint f of any input (including those independent of predictive inputs) or even the predictive component output p. This versatile framework exhibits minimal computational overhead without increasing runtime complexity. The approach thresholds an output rather than imposing an invariance or the zeroing of a derivative, but this constraint class is relevant to many physical or nonnatural systems. This provides a simple means for emulating many systems when paired with universally approximating predictive models or for integrating data-driven parameterizations into larger models while enforcing constraints like physics or conservation laws. Absolute boundaries also enhance stability in time stepping by keeping outputs realistic.

Many code packages and languages cannot compute gradients of logical branches or minimum/maximum functions. This structure bypasses that limitation, allowing constraints to be enforced during training, even on legacy systems with minimal functionality. This mitigates the need to penalize the loss function, which can impede the learning of the main objective (Rahimi et al. 2023). Enforcing constraints during training enables gradients and weight updates better suited for predicting values within the boundaries. Unlike other penalty-free approaches like projecting outputs into constrained spaces, this method does not require constant or predefined thresholds and can explicitly predict boundary values rather than asymptotically approach them, which can inhibit learning by reducing gradient magnitudes (LeCun et al. 2012). Constraints under this construction can be analytically defined, or even parameterized and learned from data during training, even simultaneously with the predictive component (e.g., a network for prediction and a network for boundary values, trained simultaneously), predicting both thresholds and values for entirely data-driven modeling. These functional forms can also be combined and stacked in larger networks or applied to nonnetwork models. Figure A1 depicts the threshold constraint layers enveloping the predictive component from section 2b, but they could be placed inside larger networks, layered, or stacked as part of a larger predictive model or applied on any nonnetwork model.

The primary benefit of this framework is enabling guaranteed prescribed bounds for trainable models during the training process. However, by representing the "if over, then change" condition as a vectorized algebraic operation rather than logical branches, it eliminates branch misprediction in compilers and maximizes scalability for broadcasting and parallelization at large scale. This has performance implications for CPUs (see Table C4) but especially for GPUs with different (or no) protocols for branching (Pharr and Fernando 2005). When inputs for both p and f are similar, constraints can be implemented with one skip connection, streamlining the architecture. These constraint structures could be equivalently

expressed using Maxout (Goodfellow et al. 2013) or nested networks for a given constraint, but maintaining a single-network form with one skip connection results in faster training and sufficient variety in constraint expression.

In the bigger picture, the framework's adaptability and guarantee of constraints make it a strong candidate for emulating systems where complex processes defy full analytical modeling but are bound by defined limits. This versatility opens doors for usage even beyond snow modeling. For instance, in two-sided threshold scenarios, M can be viewed as interpolating between two boundaries, which we anticipate as offering utility in areas like predicting drag between turbulent and viscous limits or transport in superdiffusive regimes. The approach holds the potential to contribute to a comprehensive understanding of global dynamics, opening exciting avenues for future investigation.

APPENDIX B

Data Methods

Snow data come from ground stations, aerial lidar surveys, or satellite images calibrated with ground measurements (Smyth et al. 2020). Remote sensing offers broader coverage crucial for global modeling but can suffer obfuscation and lack the resolution to capture localized snow processes. Ground sensors can capture local effects but face challenges like malfunctions, terrain biases, and limited coverage, and aerial surveys track fewer variables. The SNOTEL network, with over 900 sites, uses automated sensors like snow pillows for SWE, ultrasonic sensors for depth, and weather instruments, transmitting data via ionospheric radio reflection (National Water and Climate Center 2023).

The SNOTEL network is rare in providing simultaneous collocated hydrometeorological data across varied climates but faces notable data uncertainties like precipitation gauge undercatch or unphysical sensor values (Rasmussen et al. 2012; Hill et al. 2019). Not all sites measure all variables (Raleigh et al. 2016), with many in wind-shielded or flat terrain, limiting climate diversity for "universal" model calibration. All sites measure SWE, z, and precipitation, but the availability of variables like T_{snow}, T_{soil}, R, φ, and v varies. Removing T_{soil} [due to its correlation with $\text{ReLU}(T_{\text{air}})$] and T_{snow} increased the number of usable sites to 44 in the continental United States and 7 in Alaska after postprocessing.

Established snow science evaluation sites like Col de Porte (Lejeune et al. 2019), Kühtai (Krajči et al. 2017), Reynolds Mountain East (Reba et al. 2011), and Sodankyla (Essery et al. 2016) were tested alongside Alaska SNOTEL data, which biases the evaluation toward lower elevations and similar $\overline{z}_+$. To address this, data from Yala Basecamp (Stigter et al. 2021; Shea et al. 2015) and two Rofental sites (Warscher et al. 2024) were included, though additional daily observational time series for all variables are presently minimal. Noncollocated measurements of z, SWE, or snow density ρ_{snow} can lead to inconsistencies and biases in bulk density estimates. Training on such inconsistencies could benefit accuracy in large-scale climate modeling with gridded data, as it better matches the anticipated variability from coarse graining, but it can also introduce validation biases if observation protocols differ from the training data. Therefore, all training and testing sites were assessed for the collocation of z and SWE data collection using literature, imagery like Smyth et al. (2020), or direct communication with site representatives. This limited density evaluations to Kühtai and CONUS/Alaskan SNOTEL sites.

Phenomena like wind drift can confound snow pillow data, as strong winds away from the wind sensor can push snow onto the pillow, creating positive measured dz/dt with no precipitation and insufficient wind speeds for drift, as observed by Meyer et al. (2012) in some SNOTEL sites. Our constraints in section 2b can be violated by such events. However, only 1.5% of the training data and 2.2% of the testing data showed $dz/dt > 0$ without precipitation, and established models like Crocus and SNOWPACK also do not account for positive growth due to wind redistribution when run standalone. These models instead focus on the compaction or erosion of snow by wind (Vionnet et al. 2012; Lehning et al. 2002), destructive effects our model captures. Ongoing work aims to integrate redistribution effects into snow models, and as this paper serves to introduce a framework compared to existing parameterizations, we find accounting for such effects beyond the scope of this study, representing an exciting area for future research. We invite interested parties to revise the variables, thresholds, and datasets in this initial formulation to the benefit of the community, as we envision broader adoption of this proposal beyond our specific formulation and calibration.

a. Data cleaning procedures

Established methods exist for cleaning SNOTEL daily SWE, T_{air}, and precipitation values, beginning with Serreze et al. (1999) and extended by Yan et al. (2018), but little consensus exists for depth or meteorological variables, particularly those available at hourly frequencies.

Raw hourly and daily time series for all input variables were retrieved from the NRCS database, covering all entries available up to 2 February 2024. Bounds were applied to each sensor based on physical limits and limits in the SNOTEL sensor handbook (USDA 2010), removing any violative data. All solar, humidity, and wind speed time series were manually inspected, and any suspect or unphysical periods were flagged [see the function `manual_filter()` in the code repository for a list of suspect periods]. The following steps were then taken in order per site:

- Weekly maxima of hourly wind speeds were determined, from which the median $\tilde{w}_{\text{max}}$ and interquartile range (IQR) were calculated. Hourly wind values w_i were flagged if $(w_i - \tilde{w}_{\text{max}})/\text{IQR} > 6$. For time series with over 24 flagged values, "blocks" of flagged values were grown in 72-h steps until no further flags existed. If more than 5% of values in a block were flagged, all observations in that block were flagged.
- Unflagged hourly wind speed, solar radiation, relative humidity, and air temperature observations were binned into

2-week windows and by hour, with the mean calculated for each hourly bin to generate an annual profile (24 h per 26 biweeks). Gaps of 6 h or less were filled using linear interpolation, while gaps of 6–24 h were filled using the appropriate profile.

- For hourly z data, for each nonmissing/flagged observation z_i, the values $dz_+ = z_{i+1} - z_i$, $dz_- = z_i - z_{i-1}$, and $dt_+ = t_{z_{i+1}} - t_{z_i}$ were calculated. A threshold $Z = 20$ in. was picked, and z_i values having $dz_- \geq Z$ and $dz_+ \leq -Z$, or $dz_- \leq -Z$ and $dz_+ \geq -Z$ were flagged. A "rut" of bad data began when $dz_- \geq Z$ and $|dz_+| \leq Z$ and continued until $|dz_+| \geq Z$, and all observations in a rut were flagged. A rut lasting more than 20 observations or $dt_+ > 30$ days resulted in all observations being flagged until $z_i = 0$. From April through August, after the first time reaching $z_i = 0$, nonzero z_i values were flagged. All z_i values having z_i/SWE_i less than 1, over 50, or missing $\mathrm{SWE_i}$ were flagged. This procedure was iterated until no more z_i values were flagged and was designed based on the structure of sensor errors in the hourly z data. This was only for the hourly data, as daily SNOTEL depth data are quality checked.
- All hourly time series were then binned into three 8-h bins per day. For solar radiation, wind speed, air temperature, and humidity, the mean of all nonmissing/flagged values per bin was determined (if all were missing, a missing value was given). These three averages were averaged to create the daily values, and a daily value was reported as missing if any 8-h bin average was missing. For hourly z, the day's value was the first available observation, or a missing value if all observations that day were missing.
- The annual maxima of all generated daily solar radiation observations were determined, and the median $\tilde{R}_{\mathrm{max}}$ and IQR of these maxima were determined, and all daily solar values R_i with a score $(R_i - \tilde{R}_{\mathrm{max}})/\mathrm{IQR} > 2$ were excised. This only removed a handful of individual irregular spikes in the rolled-up data and left most sites unaltered.
- The converted daily time series from hourly data was coalesced with the raw daily time series to form a complete daily time series for the site. Raw daily z and air temperature values took priority over converted-hourly values if both values existed, and converted-hourly values for solar radiation, relative humidity, and wind speed took priority over daily values if both existed (only the raw daily data were used for SWE and accumulated precipitation).
- Air temperature corrections to the air temperature data were applied in accordance with Atwood et al. (2023) and associated metadata of which sites to correct, as of May 2024.
- Standard quality control procedures and flagging for SWE, accumulated precipitation values, and air temperature as given in Serreze et al. (1999) and extended by Yan et al. (2018) were implemented. Inconsistent water years with maximum SWE at least 5% greater than the associated accumulated precipitation value were excised in accordance with this protocol.
- The remaining accumulated precipitation observations were edited to account for gauge undercatch, following the procedure outlined in Livneh et al. (2014) used in Yan et al. (2019).
- Daily time series of z were compared to the quality-controlled SWE, and any values showing z_i/SWE_i less than 1, over 50, or missing SWE_i were excised.
- For all variables, gaps of 3 days or less were then filled via linear interpolation.
- All data were then scaled into SI units (z, SWE, and accumulated precipitation to meters from inches, relative humidity scaled from 0 to 1, and wind speed to meters per second from kilometers per hour).
- Only days with complete cases (no missing values in all variables) were extracted, and sequential differences $\Delta z_i = z_{i+1} - z_i$, $\Delta\mathrm{SWE}_i = \mathrm{SWE}_{i+1} - \mathrm{SWE}_i$, and $p_i = \mathrm{AP}_{i+1} - \mathrm{AP}_i$ were calculated, where AP is accumulated precipitation. Only values where $\Delta t_i = t_{i+1} - t_i = 1$ day were kept, and the target $(dz/dt)_i \approx \Delta z_i/\Delta t_i$ was created, as well as analogous $d\mathrm{SWE}/dt_i$ and $P_i = p_i/\Delta t_i$. Days with $P_i < 0$ due to resetting of the water year were changed to $P_i = 0$.

The data at this point were saved. Upon importing for usage in model training, precipitation P_i was split into snow P_{snow} and rain P_{rain} based off temperature and humidity values using the snow fraction equation from Jennings et al. (2018) with over an 88% success rate across the Northern Hemisphere:

$$f_{\mathrm{snow}} = \frac{1}{1 + e^{\alpha + \beta T_{\mathrm{air}} + \gamma\varphi}}, \tag{B1}$$

with $\alpha = -10.04$, $\beta = 1.41^{\circ}\mathrm{C}^{-1}$, and $\gamma = 9$ (with the relative humidity $\varphi \in [0, 1]$). The correlation of P_{rain} data after quality control procedures and feature engineering to dz/dt and $d\mathrm{SWE}/dt$ was the lowest of all variables among the training data at $r = -0.016$ and 0.005, respectively, informing the choice to also remove it from the dataset to further increase computational simplicity and scalability of the final model. Reintroduction and retraining including the rain variable on the optimized model structure to verify this choice resulted in negligible changes in performance at the site level and on average.

Three exceptions apply to the above procedure specifically for this set of training data, which might not apply for other SNOTEL sites, and are as follows:

- Air temperatures in the continental United States were bounded below by −40°C, and air temperatures in Alaska were bounded below by −50°C instead of the instrument limit of −60°C to remove on average 1–2 individual suspect temperature spikes per site.
- All raw z_i values > 175 in. were flagged (a bound solely for removing unphysical sensor spikes in this specific training data and should be checked for alternative data).
- For SNOTEL site 1122, the averaged air temperature hourly time series took priority over the daily time series when coalescing data.

The code to scrape SNOTEL data from the NRCS database and apply the above processing as well as a tutorial has been made publicly available in the code repository.

The other sites were provided varying degrees of quality control beyond unit conversion and generation of targets

dz/dt and $d\text{SWE}/dt$ from the resulting data. For Col de Porte, Kühtai, and Reynolds Mountain East, no action was taken beyond collecting data into daily time series (averaging all data except for using the first available z, SWE measurement per day, no need for intermediate 8-h blocks). Precipitation was also split into rain and snow following the same procedure as SNOTEL data for Kühtai. In the Rofental's Bella Vista site data, a nonzero offset of precipitation data was subtracted from dates after 1 January 2022, and otherwise, both sites were collected to daily data directly and treated with the same precipitation undercatch procedures as the SNOTEL data. At the Yala Basecamp site, only the 2018 year was taken due to feature availability. All negative SWE and z values were set to zero, and gaps in the SWE data up to nine consecutive values were filled with linear interpolation. The same undercatch procedures as SNOTEL were provided, and data were aggregated to the day level directly (for z, the median was taken to ignore sensor spikes, otherwise the mean for all variables). For Sodankyla, the following measures were taken from the raw data series beyond unit conversion, target feature creation in the same manner as the SNOTEL sites, and direct aggregation to daily level:

- 1-min SWE data were aggregated to the 10-min level (the same level as other variables).
- All 10-min variables had gaps up to nine consecutive values filled with linear interpolation.
- All solar radiation data less than zero were set to zero.
- Missing z data from May to November 2016 were set to zero.
- The same undercatch procedure was applied to precipitation data as the SNOTEL data.

APPENDIX C

Hyperparameters and Model Benchmarking

a. Hyperparameters

Optimal hyperparameters are summarized in Table C1. Scores were evaluated using 44-fold leave-one-out cross validation with a batch size of 64, tracking performance every 10 epochs over 200 epochs. Most time-series trials exhibited optimal performance when training for approximately 100 epochs. A nonzero value of n_2 emphasized extreme points in the custom loss function, enhancing accumulated predictions, particularly for datasets with few extreme samples. The optimal hyperparameters yielded a network size of 435 trainable parameters for the z network and 540 for the SWE network, compared to around 50 empirically tuned constants (parameters and internal code) in Snow17 for predicting SWE and z.

b. Model benchmarking

Table C2 presents the means and medians of different model configurations for time-series generation at the testing sites. The neural parameterization is also compared to another network M_{ifelse}, which has an identical predictive structure but has no boundaries during training and only calculates and applies thresholds posttraining through vectorized if/else logic. The regression RMSE scores between M and M_{ifelse} on training and testing data were within 1–2 mm day^{-1}, with M_{ifelse} performing better on training regression, despite notable superiority in M for time-series accuracy. This trend of slightly better training regression for M_{ifelse} but worse time-series generation persisted across repeated training trials, reinforcing the notion that incorporating bounds during training enhances physical representation and generalizability, rather than minimizing training loss at the expense of other beneficial properties.

Table C3 lists the p values from the Wilcoxon signed-rank tests in this study. This nonparametric test assesses the significance of differences between matched samples (time series RMSE, in this case), comparing the performance of the presented framework over Snow17 and M_{ifelse}. The lack of significance for M at validation sites was expected, as all models were calibrated for performance on these data. However, significant improvements in out-of-sample testing data highlight the advantages of the presented approach. The significant difference of the full model $\tilde{M}$ from Snow17 for validation sites was unexpected, potentially due to Snow17's calibration prioritizing SWE directly (one variable), while $\tilde{M}$ benefits from optimizing z and SWE as direct inputs for SWE prediction.

Table C4 presents time/memory benchmarking. Testing was conducted on one Intel i9 CPU (no GPU). Models were tested in "Column" mode, processing one location's inputs at a time (like a site simulation), and in "1.5M Grid" and "15M Grid" modes, evaluating input vectors of ~1.5 million and ~15 million inputs at once (like a global land model at 10 or ~3 km land resolution), by stacking 14 or 141 copies of all SNOTEL inputs, respectively. Average memory/time for a single evaluation (excluding garbage collection and compilation) was tracked and normalized by the number of

TABLE C1. Hyperparameter results. For time series, NSE and RMSE were the primary metrics, while RMSE was the main metric for regression. The term $n_1 = 2$ (L2-like metric) provided the lowest RMSE for all choices, which is unsurprising. However, it is interesting that $n_1 = 2$ also minimized SPE (an L1-like metric) compared to $n_1 = 1$.

Parameter	Description	Range	Series score: z	Series score: SWE	Regression score: dz/dt
N	Averaging consecutive N days	1, 2, 3	1	1	1
n	Width of mixing layer	3, 4, 5, 6	4	5	6
n_1	Power scaling of prediction error	1, 2	2	2	2
n_2	Power scaling of target magnitude	0, 0.1, 1, 2, 4	4	2	2

TABLE C2. Performance of the models (for SWE and z) and parameterizations (for z) in this study across testing sites, using labels from section 2f. Medians are listed, with the mean in parentheses alongside the median. Another network M_{ifelse} has been included, which trains without constraints and applies them during testing with if/else statements, to highlight the performance gains over out-of-sample data from including the framework developed in this paper. Subscripts on metrics indicate whether depth or SWE was benchmarked.

	Standalone model (SWE and z)		Parameterization (z)		
Metric	$\tilde{M}$	SN17	M	SN17O	M_{ifelse}
MAE_z (cm)	6.9 (7.8)	8.2 (16.1)	4.6 (5.9)	6.9 (8.9)	6.4 (8.0)
RMSE_z (cm)	11.3 (12.9)	12.9 (23.7)	8.3 (9.6)	10.1 (14.3)	10.6 (13.0)
NSE_z	0.915 (0.870)	0.914 (0.347)	0.955 (0.936)	0.937 (0.782)	0.925 (0.875)
SPE_z (%)	11.9 (13.4)	11.5 (27.6)	8.8 (9.6)	9.8 (15.4)	11.5 (13.5)
MAE_{SWE} (cm)	2.1 (2.3)	2.5 (4.2)	—	—	—
RMSE_{SWE} (cm)	3.7 (4.1)	4.4 (7.4)	—	—	—
NSE_{SWE}	0.931 (0.862)	0.933 (0.663)	—	—	—
SPE_{SWE} (%)	11.3 (14.2)	11.3 (21.4)	—	—	—

TABLE C3. Statistical significance testing of model RMSE at validation and testing sites. The p values of the Wilcoxon signed-rank tests are shown, which compares the predictive power of two models. Labels follow those in Table C2. Significant ($p < 0.05$) values are highlighted in bold. The term M_{ifelse} has been added to underscore the significant improvement on out-of-sample data when applying the demonstrated framework.

Site RMSE	M vs SN17O (z)	$\tilde{M}$ vs SN17 (z)	$\tilde{M}$ vs SN17 (SWE)	M vs M_{ifelse} (z)
Validation sites	0.673	**0.006**	**0.029**	0.903
Testing sites	**0.013**	0.135	0.268	**0.0003**

TABLE C4. Time and memory benchmarking of all models (for z and SWE) and parameterizations (for z), listing required resources per evaluation for single instances (Column) or per instance over roughly 1.5 million or 15 million instances (1.5M Grid and 15M Grid) simultaneously. Snow17 only evaluates single instances, and M_{ifelse} is listed to compare constraint layers against vectorized if/else postprocessing. All benchmarks were evaluated on a single Intel i9 CPU.

	Standalone model (z and SWE)		Parameterization (z)		
Metric	$\tilde{M}$	SN17	M	SN17O	M_{ifelse}
Column, T (μs)	2.6	3.6	1.3	3.8	1.0
Column, allocated memory (KB)	2.5	1.6	1.2	1.7	1.1
1.5M Grid, T (μs)	0.21	—	0.099	—	0.11
1.5M Grid, allocated memory (KB)	0.80	—	0.37	—	0.34
15M Grid, T (μs)	0.21	—	0.099	—	0.11
15M Grid, allocated memory (KB)	0.80	—	0.37	—	0.34

inputs. Column mode was averaged over 10 SNOTEL data passes (1.06 million trials), while 1.5M and 15M Grid results were averaged over 250 trials. Snow17 can only iterate between locations in a Column-like mode. Both M and M_{ifelse} determine boundaries and adjust outputs accordingly, but M does this structurally, while M_{ifelse} uses vectorized boundary creation and broadcasts conditional if/else logic, creating the opportunity for branch misprediction effects.

While M_{ifelse} is slightly faster with less memory in Column mode (comparing and adjusting one value against two values is quicker than processing three values through matrix multiplication), M is faster by about 10% over gridded inputs (0.147 and 1.47 s total for all 1.5 million and 15 million inputs, respectively). Snow17 with data assimilation requires more time and memory than without, as assimilation occurs after the primary evaluation. The model $\tilde{M}$ is larger than standalone Snow17 (and about twice that of M, as expected) but evaluates faster, which benefits long simulations across many locations in global models. However, M is quicker and requires less memory than Snow17. Results may vary based on programming languages, libraries, compilers, or protocols used for implementation, but one could always train utilizing the framework (the main benefit) and alter boundary implementation afterward as desired. We anticipate these findings would extend to GPUs that often lack optimized branching, suggesting further advantages for large-scale model integration. Additional comprehensive benchmarking remains an avenue for future research.

REFERENCES

Anderson, E., 2006: Snow accumulation and ablation model – SNOW-17. Tech. Doc., 61 pp., https://www.weather.gov/media/owp/oh/hrl/docs/22snow17.pdf.

Anderson, E. A., 1976: A point energy and mass balance model of a snow cover. NOAA Tech. Rep. NWS 19, 172 pp., https://repository.library.noaa.gov/view/noaa/6392.

Apley, D. W., and J. Zhu, 2020: Visualizing the effects of predictor variables in black box supervised learning models. *J. Roy. Stat. Soc.*, **82B**, 1059–1086, https://doi.org/10.1111/rssb.12377.

Atwood, J., and Coauthors, 2023: Evaluation of YSI temperature correction equations for bias-reducing SNOTEL network temperature data. 11 pp., https://www.nrcs.usda.gov/sites/default/files/2023-05/Final_Temperature_Correction_Study05262023.pdf.

Bair, E. H., A. Abreu Calfa, K. Rittger, and J. Dozier, 2018: Using machine learning for real-time estimates of snow water equivalent in the watersheds of Afghanistan. *Cryosphere*, **12**, 1579–1594, https://doi.org/10.5194/tc-12-1579-2018.

Ban, N., and Coauthors, 2021: The first multi-model ensemble of regional climate simulations at kilometer-scale resolution, part I: Evaluation of precipitation. *Climate Dyn.*, **57**, 275–302, https://doi.org/10.1007/s00382-021-05708-w.

Beucler, T., M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, 2021: Enforcing analytic constraints in neural networks emulating physical systems. *Phys. Rev. Lett.*, **126**, 098302, https://doi.org/10.1103/PhysRevLett.126.098302.

Boone, A., and P. Etchevers, 2001: An Intercomparison of three snow schemes of varying complexity coupled to the same land surface model: Local-scale evaluation at an alpine site. *J. Hydrometeor.*, **2**, 374–394, https://doi.org/10.1175/1525-7541(2001)002<0374:AIOTSS>2.0.CO;2.

Bormann, K. J., S. Westra, J. P. Evans, and M. F. McCabe, 2013: Spatial and temporal variability in seasonal snow density. *J. Hydrol.*, **484**, 63–73, https://doi.org/10.1016/j.jhydrol.2013.01.032.

Bruland, O., A. Faerevag, I. Steinsland, G. E. Liston, and K. Sand, 2015: Weather SDM: Estimating snow density with high precision using snow depth and local climate. *Hydrol. Res.*, **46**, 494–506, https://doi.org/10.2166/nh.2015.059.

Brun, E., E. Martin, V. Simon, C. Gendre, and C. Coleou, 1989: An energy and mass model of snow cover suitable for operational avalanche forecasting. *J. Glaciol.*, **35**, 333–342, https://doi.org/10.3189/S0022143000009254.

——, V. Vionnet, A. Boone, B. Decharme, Y. Peings, R. Valette, F. Karbou, and S. Morin, 2013: Simulation of northern Eurasian local snow depth, mass, and density using a detailed snowpack model and meteorological reanalyses. *J. Hydrometeor.*, **14**, 203–219, https://doi.org/10.1175/JHM-D-12-012.1.

Chen, R. T. Q., Y. Rubanova, J. Bettencourt, and D. Duvenaud, 2018: Neural ordinary differential equations. arXiv, 1806.07366v5, https://doi.org/10.48550/ARXIV.1806.07366.

Clark, M. P., and Coauthors, 2015: Improving the representation of hydrologic processes in Earth System Models. *Water Resour. Res.*, **51**, 5929–5956, https://doi.org/10.1002/2015WR017096.

Cui, G., M. Anderson, and R. Bales, 2023: Mapping of snow water equivalent by a deep-learning model assimilating snow observations. *J. Hydrol.*, **616**, 128835, https://doi.org/10.1016/j.jhydrol.2022.128835.

De Michele, C., F. Avanzi, A. Ghezzi, and C. Jommi, 2013: Investigating the dynamics of bulk snow density in dry and wet conditions using a one-dimensional model. *Cryosphere*, **7**, 433–444, https://doi.org/10.5194/tc-7-433-2013.

Diro, G. T., and L. Sushama, 2018: Snow–precipitation coupling and related atmospheric feedbacks over North America. *Atmos. Sci. Lett.*, **19**, e831, https://doi.org/10.1002/asl.831.

Dong, S., and N. Ni, 2021: A method for representing periodic functions and enforcing exactly periodic boundary conditions with deep neural networks. *J. Comput. Phys.*, **435**, 110242, https://doi.org/10.1016/j.jcp.2021.110242.

Duan, S., P. Ullrich, M. Risser, and A. Rhoades, 2024: Using temporal deep learning models to estimate daily snow water equivalent over the Rocky Mountains. *Water Resour. Res.*, **60**, e2023WR035009, https://doi.org/10.1029/2023WR035009.

Dutra, E., G. Balsamo, P. Viterbo, P. M. A. Miranda, A. Beljaars, C. Schär, and K. Elder, 2010: An improved snow scheme for the ECMWF land surface model: Description and offline validation. *J. Hydrometeor.*, **11**, 899–916, https://doi.org/10.1175/2010JHM1249.1.

Ebner, P. P., and Coauthors, 2021: Evaluating a prediction system for snow management. *Cryosphere*, **15**, 3949–3973, https://doi.org/10.5194/tc-15-3949-2021.

Essery, R., A. Kontu, J. Lemmetyinen, M. Dumont, and C. B. Ménard, 2016: A 7-year dataset for driving and evaluating snow models at an Arctic site (Sodankylä, Finland). *Geosci. Instrum. Methods Data Syst.*, **5**, 219–227, https://doi.org/10.5194/gi-5-219-2016.

Fontrodona-Bach, A., B. Schaefli, R. Woods, A. J. Teuling, and J. R. Larsen, 2023: NH-SWE: Northern Hemisphere snow water equivalent dataset based on in situ snow depth time series. *Earth Syst. Sci. Data*, **15**, 2577–2599, https://doi.org/10.5194/essd-15-2577-2023.

Gao, L., L. Zhang, Y. Shen, Y. Zhang, M. Ai, and W. Zhang, 2021: Modeling snow depth and snow water equivalent distribution and variation characteristics in the Irtysh River Basin, China. *Appl. Sci.*, **11**, 8365, https://doi.org/10.3390/app11188365.

Goodfellow, I. J., D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, 2013: Maxout networks. arXiv, 1302.4389v4, https://doi.org/10.48550/ARXIV.1302.4389.

Hedrick, A. R., and Coauthors, 2018: Direct insertion of NASA Airborne Snow Observatory-derived snow depth time series into the *iSnobal* energy balance snow model. *Water Resour. Res.*, **54**, 8045–8063, https://doi.org/10.1029/2018WR023190.

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.

Hinton, G., N. Srivastava, and K. Swersky, 2012: Lecture 6e: Rmsprop: Divide the gradient by a running average of its recent magnitude. University of Toronto, https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Innes, M., 2018: Flux: Elegant machine learning with Julia. *J. Open Source Software*, **3**, 602, https://doi.org/10.21105/joss.00602.

——, and Coauthors, 2018: Fashionable modelling with flux. arXiv, 1811.01457v3, https://doi.org/10.48550/arXiv.1811.01457.

Jennings, K. S., T. S. Winchell, B. Livneh, and N. P. Molotch, 2018: Spatial variation of the rain-snow temperature threshold across the Northern Hemisphere. *Nat. Commun.*, **9**, 1148, https://doi.org/10.1038/s41467-018-03629-7.

Jiang, C. M., K. Kashinath, Prabhat, and P. Marcus, 2020: Enforcing physical constraints in CNNs through differentiable PDE layer. *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, Online, ICLR, 2–6, https://openreview.net/forum?id=q2noHUqMkK.

Kapnick, S. B., and Coauthors, 2018: Potential for western US seasonal snowpack prediction. *Proc. Natl. Acad. Sci. USA*, **115**, 1180–1185, https://doi.org/10.1073/pnas.1716760115.

Kojima, K., 1967: Densification of seasonal snow cover. *Phys. Snow Ice: Proc.*, **1**, 929–952.

Kouki, K., P. Räisänen, K. Luojus, A. Luomaranta, and A. Riihelä, 2022: Evaluation of Northern Hemisphere snow water equivalent in CMIP6 models during 1982–2014. *Cryosphere*, **16**, 1007–1030, https://doi.org/10.5194/tc-16-1007-2022.

Krajči, P., R. Kirnbauer, J. Parajka, J. Schöber, and G. Blöschl, 2017: The Kühtai data set: 25 years of lysimetric, snow pillow, and meteorological measurements. *Water Resour. Res.*, **53**, 5158–5165, https://doi.org/10.1002/2017WR020445.

Lawrence, D. M., and Coauthors, 2019: The Community Land Model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. *J. Adv. Model. Earth Syst.*, **11**, 4245–4287, https://doi.org/10.1029/2018MS001583.

LeCun, Y. A., L. Bottou, G. B. Orr, and K.-R. Müller, 2012: Efficient BackProp. *Neural Networks: Tricks of the Trade*, G. Montavon, G. B. Orr and K. R. Müller, Eds., Lecture Notes in Computer Science, Vol. 7700, Springer, 9–48, https://doi.org/10.1007/978-3-642-35289-8_3.

Lehning, M., P. Bartelt, B. Brown, C. Fierz, and P. Satyawali, 2002: A physical SNOWPACK model for the Swiss avalanche warning: Part II. Snow microstructure. *Cold Reg. Sci. Technol.*, **35**, 147–167, https://doi.org/10.1016/S0165-232X(02)00073-3.

Lejeune, Y., M. Dumont, J.-M. Panel, M. Lafaysse, P. Lapalus, E. Le Gac, B. Lesaffre, and S. Morin, 2019: 57 years (1960–2017) of snow and meteorological observations from a mid-altitude mountain site (Col de Porte, France, 1325 m of altitude). *Earth Syst. Sci. Data*, **11**, 71–88, https://doi.org/10.5194/essd-11-71-2019.

Liu, J., R. Chen, S. Ma, C. Han, Y. Ding, S. Guo, and X. Wang, 2024: The challenge of monitoring snow surface sublimation in winter could be resolved with structure-from-motion photogrammetry. *J. Hydrol.*, **630**, 130733, https://doi.org/10.1016/j.jhydrol.2024.130733.

Livneh, B., J. S. Deems, D. Schneider, J. J. Barsugli, and N. P. Molotch, 2014: Filling in the gaps: Inferring spatially distributed precipitation from gauge observations over complex terrain. *Water Resour. Res.*, **50**, 8589–8610, https://doi.org/10.1002/2014WR015442.

Luijting, H., D. Vikhamar-Schuler, T. Aspelien, Å. Bakketun, and M. Homleid, 2018: Forcing the SURFEX/Crocus snow model with combined hourly meteorological forecasts and gridded observations in southern Norway. *Cryosphere*, **12**, 2123–2145, https://doi.org/10.5194/tc-12-2123-2018.

Lundy, C. C., R. L. Brown, E. E. Adams, K. W. Birkeland, and M. Lehning, 2001: A statistical validation of the snowpack model in a Montana climate. *Cold Reg. Sci. Technol.*, **33**, 237–246, https://doi.org/10.1016/S0165-232X(01)00038-6.

Marks, D., S. Havens, M. Johnson, and M. Sandusky, 2018: Pysnobal, v2.4.1. Zenodo, https://doi.org/10.5281/ZENODO.1301290.

Meloche, J., A. Langlois, N. Rutter, D. McLennan, A. Royer, P. Billecocq, and S. Ponomarenko, 2022: High-resolution snow depth prediction using Random Forest algorithm with topographic parameters: A case study in the Greiner watershed, Nunavut. *Hydrol. Processes*, **36**, e14546, https://doi.org/10.1002/hyp.14546.

Menard, C. B., and Coauthors, 2021: Scientific and human errors in a snow model intercomparison. *Bull. Amer. Meteor. Soc.*, **102**, E61–E79, https://doi.org/10.1175/BAMS-D-19-0329.1.

Meyer, J., J. Horel, P. Kormos, A. Hedrick, E. Trujillo, and S. M. Skiles, 2023: Operational water forecast ability of the HRRR-iSnobal combination: An evaluation to adapt into production environments. *Geosci. Model Dev.*, **16**, 233–250, https://doi.org/10.5194/gmd-16-233-2023.

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

Molnar, C., 2022: *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. Christoph Molnar, 317 pp., https://christophm.github.io/interpretable-ml-book.

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

National Water and Climate Center, 2023: Telemetry and data transmission. USDA National Resources Conservation Service, https://www.nrcs.usda.gov/wps/portal/wcc/home/aboutUs/monitoringPrograms/telemetry/.

Niu, G.-Y., and Coauthors, 2011: The community Noah Land Surface Model with Multiparameterization Options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, https://doi.org/10.1029/2010JD015139.

Pharr, M., and R. Fernando, 2005: *GPU Gems 2: Programming Techniques for High-Performance Graphics and General-Purpose Computation.* Addison-Wesley Professional, 814 pp.

Rahimi, I., A. H. Gandomi, F. Chen, and E. Mezura-Montes, 2023: A Review on constraint handling techniques for population-based algorithms: From single-objective to multi-objective optimization. *Arch. Comput. Methods Eng.*, **30**, 2181–2209, https://doi.org/10.1007/s11831-022-09859-9.

Raleigh, M. S., B. Livneh, K. Lapo, and J. D. Lundquist, 2016: How does availability of meteorological forcing data impact physically based snowpack simulations? *J. Hydrometeor.*, **17**, 99–120, https://doi.org/10.1175/JHM-D-14-0235.1.

Rasmussen, R., and Coauthors, 2012: How well are we measuring snow: The NOAA/FAA/NCAR winter precipitation test bed. *Bull. Amer. Meteor. Soc.*, **93**, 811–829, https://doi.org/10.1175/BAMS-D-11-00052.1.

Reba, M. L., D. Marks, M. Seyfried, A. Winstral, M. Kumar, and G. Flerchinger, 2011: A long-term data set for hydrologic modeling in a snow-dominated mountain catchment. *Water Resour. Res.*, **47**, W07702, https://doi.org/10.1029/2010WR010030.

Schär, C., and Coauthors, 2020: Kilometer-scale climate models: Prospects and challenges. *Bull. Amer. Meteor. Soc.*, **101**, E567–E587, https://doi.org/10.1175/BAMS-D-18-0167.1.

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.

Shea, J., P. Wagnon, W. Immerzeel, R. Biron, F. Brun, and F. Pellicciotti, 2015: A comparative high-altitude meteorological analysis from three catchments in the Nepalese Himalaya. *Int. J. Water Resour. Dev.*, **31**, 174–200, https://doi.org/10.1080/07900627.2015.1020417.

Smyth, E. J., M. S. Raleigh, and E. E. Small, 2020: Improving SWE estimation with data assimilation: The influence of snow depth observation timing and uncertainty. *Water*

Resour. Res., **56**, e2019WR026853, https://doi.org/10.1029/2019WR026853.

Song, Y., W.-P. Tsai, J. Gluck, A. Rhoades, C. Zarzycki, R. McCrary, K. Lawson, and C. Shen, 2024: LSTM-based data integration to improve snow water equivalent prediction and diagnose error sources. *J. Hydrometeor.*, **25**, 223–237, https://doi.org/10.1175/JHM-D-22-0220.1.

Spehlmann, K. A., E. S. Euskirchen, and S. L. Stuefer, 2025: Sublimation measurements of tundra and taiga snowpack in Alaska. *Cryosphere*, **19**, 1739–1755, https://doi.org/10.5194/tc-19-1739-2025.

Steele, H., E. E. Small, and M. S. Raleigh, 2024: Demonstrating a hybrid machine learning approach for snow characteristic estimation throughout the western United States. *Water Resour. Res.*, **60**, e2023WR035805, https://doi.org/10.1029/2023WR035805.

Stigter, E. E., J. F. Steiner, I. Koch, T. M. Saloranta, J. D. Kirkham, and W. W. Immerzeel, 2021: Energy and mass balance dynamics of the seasonal snowpack at two high-altitude sites in the Himalaya. *Cold Reg. Sci. Technol.*, **183**, 103233, https://doi.org/10.1016/j.coldregions.2021.103233.

Tanniru, S., and R. Ramsankaran, 2023: Machine learning-based estimation of high-resolution snow depth in Alaska using passive microwave remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **16**, 6007–6025, https://doi.org/10.1109/JSTARS.2023.3287410.

USDA, 2010: Data parameter. *Part 622 Snow Survey and Water Supply Forecasting National Engineering Handbook*, USDA, 3–10, https://directives.nrcs.usda.gov//sites/default/files2/1720456630/Chapter%202%20-%20Data%20Parameters.pdf.

Vafakhah, M., A. Nasiri Khiavi, S. Janizadeh, and H. Ganjkhanlo, 2022: Evaluating different machine learning algorithms for snow water equivalent prediction. *Earth Sci. Inform.*, **15**, 2431–2445, https://doi.org/10.1007/s12145-022-00846-z.

Viallon-Galinier, L., P. Hagenmuller, and M. Lafaysse, 2020: Forcing and evaluating detailed snow cover models with stratigraphy observations. *Cold Reg. Sci. Technol.*, **180**, 103163, https://doi.org/10.1016/j.coldregions.2020.103163.

Vionnet, V., E. Brun, S. Morin, A. Boone, S. Faroux, P. Le Moigne, E. Martin, and J.-M. Willemet, 2012: The detailed snowpack scheme Crocus and its implementation in SURFEX v7.2. *Geosci. Model Dev.*, **5**, 773–791, https://doi.org/10.5194/gmd-5-773-2012.

——, and Coauthors, 2019: Sub-kilometer precipitation datasets for snowpack and glacier modeling in Alpine Terrain. *Front. Earth Sci.*, **7**, 182, https://doi.org/10.3389/feart.2019.00182.

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

Warscher, M., T. Marke, E. Rottler, and U. Strasser, 2024: Operational and experimental snow observation systems in the upper Rofental: Data from 2017 to 2023. *Earth Syst. Sci. Data*, **16**, 3579–3599, https://doi.org/10.5194/essd-16-3579-2024.

Wever, N., L. Schmid, A. Heilig, O. Eisen, C. Fierz, and M. Lehning, 2015: Verification of the multi-layer SNOWPACK model with different water transport schemes. *Cryosphere*, **9**, 2271–2293, https://doi.org/10.5194/tc-9-2271-2015.

Wilcoxon, F., 1945: Individual comparisons by ranking methods. *Biometrics Bull.*, **1**, 80–83, https://doi.org/10.2307/3001968.

Xu, L., and P. Dirmeyer, 2011: Snow-atmosphere coupling strength in a global atmospheric model. *Geophys. Res. Lett.*, **38**, L13401, https://doi.org/10.1029/2011GL048049.

Yan, H., N. Sun, M. Wigmosta, R. Skaggs, Z. Hou, and R. Leung, 2018: Next-generation intensity-duration-frequency curves for hydrologic design in snow-dominated environments. *Water Resour. Res.*, **54**, 1093–1108, https://doi.org/10.1002/2017WR021290.

——, ——, ——, ——, L. R. Leung, A. Coleman, and Z. Hou, 2019: Observed spatiotemporal changes in the mechanisms of extreme water available for runoff in the western United States. *Geophys. Res. Lett.*, **46**, 767–775, https://doi.org/10.1029/2018GL080260.

Yang, J., L. Jiang, K. Luojus, J. Pan, J. Lemmetyinen, M. Takala, and S. Wu, 2020: Snow depth estimation and historical data reconstruction over China based on a random forest machine learning approach. *Cryosphere*, **14**, 1763–1778, https://doi.org/10.5194/tc-14-1763-2020.

——, ——, J. Pan, J. Shi, S. Wu, J. Wang, and F. Pan, 2022: Comparison of machine learning-based snow depth estimates and development of a new operational retrieval algorithm over China. *Remote Sens.*, **14**, 2800, https://doi.org/10.3390/rs14122800.