https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

SWEDE v1.0: emulating dynamically downscaled snow water equivalent across the western United States depends on training data composition

Fanghe Zhao[1], Stefan Rahimi[1], Melissa Bukovsky[1,2], Karandeep Singh[3], Chad Thackeray[3], Jesse Norris[3], Alex Hall[3], Weichen Liu[1]

[1]Department of Atmospheric Science, University of Wyoming, WY, USA
[2]Haub School of Environment and Natural Resources, University of Wyoming, WY, USA
[3]Atmospheric and Oceanic Sciences, University of California, Los Angeles, CA, USA

Correspondence to: Fanghe Zhao (fzhao@uwyo.edu)

Abstract. Snowpack in the western United States serves as a critical seasonal freshwater reservoir, yet Earth System models (ESMs) lack the spatial resolution to resolve snow water equivalent (SWE) at scales relevant to water resource management. We present SWEDE (Snow Water Equivalent Downscaling Emulator), a dual-encoder Swin-Transformer U-Net that downscales coarse atmospheric output to 9 km SWE across the western United States without receiving coarse SWE as input. We assess how SWEDE's performance varies across a collection of diverse training setups. When trained on the Western U.S. Dynamically Downscaled Dataset (15 CMIP6 simulations from 14 ESMs from 1980–2100), SWEDE performs best (post-2050 MAE of 0.93 mm). In the WRF-ERA5 experiment, SWEDE achieves interannual mean SWE correlations of 0.89 to 0.99 and annual maximum SWE biases below 10% across all six subregions during 2014–2024. When trained on a single ESM, performance degrades, but we find that the choice of ESM has an impact here, with high-warming ESMs outperforming low-warming ones (post-2050 MAE of 1.50 versus 1.91 mm). Lastly, both the multi- and single-ESM training setups substantially outperform a historical reanalysis setup, which struggles to emulate conditions under future warming, illustrating the importance of exposing the emulator to future climate states during training. These findings establish that climate-aware training data composition, particularly the inclusion of warm-climate sources, is essential for deep learning downscaling systems intended for use beyond the historical record.

1 Introduction

Snowpack functions as a natural water reservoir across the mountainous western United States, accumulating cold-season precipitation and releasing meltwater during spring and summer when downstream demand is highest (Bales et al., 2006; Li et al., 2017). This delayed storage supports streamflow, irrigated agriculture, municipal water supply, hydropower production, and freshwater ecosystems, while also shaping drought resilience and flood risk (Barnett et al., 2005; Mankin et al., 2015; Mote et al., 2005). Snow Water Equivalent (SWE), which measures the liquid water stored in the snowpack, is the primary variable used to quantify this seasonal reservoir (Dozier et al., 2016; Painter et al., 2016). Because warming is projected to reduce snow accumulation, shift precipitation from snow to rain, and accelerate melt timing across much of the

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

western United States, reliable projections of SWE are essential for climate risk assessment and long-term water-resource planning (Mote et al., 2018; Siirila-Woodburn et al., 2021; Musselman et al., 2017).

However, obtaining reliable projections of SWE remains challenging. Earth System Models (ESMs) typically operate at spatial resolutions too coarse to resolve the topographic gradients, elevation-dependent temperature structure, and orographic precipitation processes that control snow accumulation and ablation in complex terrain (Lundquist et al., 2019; Wrzesien et al., 2018; Wrzesien et al., 2017). Even when large-scale climate forcing is reasonably represented, coarse grids smooth mountain barriers and lower model terrain, which biases snow-rain partitioning toward rain and underestimates local SWE extremes (Ikeda et al., 2010; Rasmussen et al., 2011). As a result, SWE fields produced directly by ESMs contain substantial biases and are generally unsuitable for regional water-resource applications without further refinement (Wrzesien et al., 2019; Pritchard et al., 2021). Dynamical downscaling with regional climate models can recover part of this missing spatial structure, but its computational cost limits the number of simulations, scenarios, and ensemble members that can be explored (Gutmann et al., 2016). Statistical downscaling is computationally efficient, yet conventional approaches often rely on assumptions of stationarity and may struggle to represent the strongly nonlinear and spatially heterogeneous processes that govern snowpack evolution in mountain environments (Wrzesien et al., 2018; Walton et al., 2015; Rasmussen et al., 2023).

Machine learning has recently emerged as a promising alternative for climate downscaling, particularly for variables such as precipitation and near-surface temperature, where deep neural networks can learn complex mappings from coarse atmospheric predictors to high-resolution surface responses at a fraction of the cost of dynamical models (Baño-Medina et al., 2020; Pan et al., 2019; Vandal et al., 2017). SWE presents a particularly demanding target for this class of methods (Clark et al., 2011). Its spatial distribution is shaped jointly by large-scale meteorological forcing and fine-scale terrain controls, and its statistics are highly intermittent, zero-inflated, and strongly seasonal (Lundquist et al., 2019; Sturm et al., 2010). Existing machine learning studies related to snow have mainly focused on local prediction, gap filling, or data fusion (Bair et al., 2018; King et al., 2020; Fang et al., 2022), whereas spatially complete downscaling of coarse ESM output into gridded high-resolution SWE over a domain as large as the western United States remains much less developed (Zakeri et al., 2025). This gap motivates a modeling framework that can simultaneously exploit coarse atmospheric information, preserve fine-scale physiographic controls, and represent the non-Gaussian character of SWE.

A further and often underexamined challenge is that downscaling itself is vulnerable to non-stationarity (Benestad, 2026; Dixon et al., 2016). Models trained only on historical conditions are typically optimized for interpolation within the climate states represented in the training data (Najafi et al., 2025). Under future warming, however, they may encounter temperature regimes, precipitation-phase transitions, and snow-climate relationships that differ systematically from those seen during training. This issue is a form of distribution shift, or out-of-distribution generalization, and it is especially consequential for snow because modest thermodynamic changes can sharply alter whether precipitation accumulates as snow or falls as rain in marginal elevation zones (Klos et al., 2014; Harpold et al., 2017). Although this problem is conceptually recognized in climate downscaling, its practical implications for deep-learning-based SWE projections remain insufficiently quantified (Rampal et al., 2024).

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

In this study, we address both the spatial-resolution and non-stationarity problems in machine-learning based SWE estimation. We develop the Snow Water Equivalent Downscaling Emulator (SWEDE), a dual-encoder Swin-Transformer U-Net that maps approximately 1-degree ESM meteorological predictors, together with 9-km static physiographic information, to obtain 9 km gridded daily SWE across the western United States. The architecture processes the coarse meteorological fields and high-resolution terrain variables through separate encoding branches suited to their respective spatial scales and correlation structures, fusing them in a shared decoder with attention-gated skip connections. Crucially, our approach diverges from traditional statistical downscaling, which typically refines a coarse-resolution version of the target variable. Instead, we perform simultaneous prediction and downscaling by providing no coarse SWE information to the model. In effect, the model learns to emulate the snow accumulation and ablation processes resolved by regional earth system model (RESM), reproducing its terrain-aware snow physics from coarse atmospheric boundary conditions at a fraction of the computational cost of traditional dynamical downscaling. We then perform a systematic training-data exercise to evaluate the consequences of training set design for model generalization to confirm under which configurations time-stationarity is bypassed.

This study tests three linked hypotheses. First, we hypothesize that a deep learning (DL) model can accurately emulate RESM-resolved snow accumulation and ablation across the western United States from coarse geophysical and high-resolution terrain alone, without receiving any coarse SWE as input. Second, we hypothesize that the performance of DL models trained exclusively on historical-era data degrades when predicting future SWE due to the problem of distribution shift. Third, we hypothesize that generalization to an unseen driving model depends on the diversity of climate states sampled during training rather than on data volume alone. The multi-ESM training set spans a wider envelope of forced and internal climate states than any single model provides. An emulator trained across multiple ESMs should therefore transfer zero-shot to a fully withheld large ensemble, reproducing both the forced trend and the ensemble spread of SWE through the 21st century.

2 Data Description and Processing Pipeline

All training and evaluation data are drawn from the Western U.S. Dynamically Downscaled Dataset (WUS-D3) dataset (Rahimi et al., 2024a), which provides high-resolution regional climate projections over the western United States produced by dynamically downscaling Coupled Model Intercomparison Project Phase 6 (CMIP6) (O'neill et al., 2016) global climate model outputs with the Weather Research and Forecasting (WRF) model (Skamarock et al., 2019). The WRF configuration used in WUS-D3 was previously optimized and validated against ERA5 reanalysis in Rahimi et al. (2022), establishing a performance benchmark under historically observed conditions (Zakeri et al., 2025). The WUS-D3 archive comprises WRF-downscaled output at 9 km horizontal resolution for 15 CMIP6 ESMs (listed with their native resolutions and modeling centers in Appendix B, Table B1), a 10-member CESM2 Large Ensemble (CESM-LE), and an ERA5-driven simulation (WRF-ERA5), spanning the period 1980-2100 for the ESM-driven runs under the SSP3-7.0 high-emission scenario and 1950-2024 for the ERA5-driven run. Prior to downscaling, the GCM boundary conditions (temperature, moisture, horizontal winds, surface pressure, and sea surface temperatures) are bias corrected against the ERA5 mean-state climatology,

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

a monthly varying adjustment that removes mean GCM biases while preserving synoptic-scale variability (Rahimi et al., 2024a). All fields are provided at daily temporal resolution. The simulation domain covers the western United States (approximately 30-50°N, 125-100°W) on a 340 × 270 grid, encompassing the full range of montane, arid, and maritime snow climates in the region. The WUS-D3 dataset has also been used for data-driven SWE downscaling over the western United States (Zakeri et al., 2025). That study reconstructs high-resolution SWE with a K-nearest-neighbor analog method that requires coarse-resolution SWE as input, so its skill degrades when analogs vanish under late-century warming. The present study instead maps atmospheric predictors directly to high-resolution SWE, enabling consistent application across the full 15-ESM WUS-D3 ensemble through 2100.

2.1 Input and Target Construction

The downscaling target here is high-resolution (9 km) SWE as simulated by WRF. Coarse-resolution atmospheric inputs are constructed by spatially averaging 10 × 10 blocks of the 340 × 270 WRF grid, yielding a 34 × 27 input grid at an effective resolution of approximately 90 km that approximates the native scale of the driving global climate models. This coarsening is applied uniformly to all dynamic atmospheric variables listed in Table S1, preserving physical consistency among the input fields while simulating the information loss inherent in coarse-resolution ESM output. Static terrain variables (Table S1) are retained at the native 9 km resolution, as they represent time-invariant surface properties that would be available at the target grid in any operational downscaling application. All inputs are standardized using z-score normalization with statistics computed from the ERA5 driven WRF simulation (hereafter WRF-ERA5) (Rahimi et al., 2022) over the training period (1950 to 2013). For static variables, the mean and standard deviation are computed spatially across the domain, whereas for dynamic variables they are computed over both space and time across the full training period. These standardization statistics are computed from the training set only and held fixed during validation and testing to prevent information leakage.

2.2 Training and Evaluation Design

Model training and evaluation follow a staged design in which each stage tests one named hypothesis. Stage 1 tests the emulation hypothesis: SWEDE is trained on the WRF-ERA5 from 1950 to 2013 and evaluated from 2014 to 2024. This stage assesses whether the network can emulate WRF-resolved snow processes under reanalysis-constrained conditions where the driving large-scale meteorology closely tracks observations. Stage 2 tests the training composition hypothesis: for each target ESM, SWEDE is trained on the remaining 14 ESMs, producing a leave-one-ESM-out cross-validation that tests generalization across the range of large-scale biases and climate sensitivities represented in the CMIP6 archive. Stage 3 tests the generalizability hypothesis: SWEDE is trained jointly on all 15 ESMs together with WRF-ERA5 and evaluated on a fully withheld 10-member dynamically downscaled ensemble from the Community Earth System Model version 2 Large Ensemble (CESM2-LE) (Kay et al., 2015; Rodgers et al., 2021). This tests whether the emulator reproduces both the forced trend and the ensemble spread of an ESM whose internal variability and forced response are independent of the training distribution.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Together, the three stages seek to determine whether models trained on historical data can reliably extrapolate to future climate states where temperature, precipitation phase, and snowpack characteristics fall outside the training distribution.

The structure of the WUS-D3 grids is central to interpreting these experiments. Each WUS-D3 simulation employs two one-way nested WRF domains: an outer 45 km domain (d01), and an inner 9 km domain (d02) covering the western United States. Under one-way nesting, d01 supplies lateral boundary conditions to d02 but receives no feedback in return, so the meteorology on the two grids evolves as separate simulations, and d01 fields are not smoothed versions of d02. Across all three stages, both the training and evaluation inputs are the synthetically coarsened 9 km d02 fields described in Section 2.1, so the reported skill reflects performance on inputs that share the spatial structure of the high-resolution target. To test sensitivity to this construction, we additionally perform an inference-only experiment in which the model trained on all 15 ESMs and ERA5 is driven by the 45 km d01 output regridded to the same approximately 90 km input grid. Because the d01 fields come from an independently evolving coarse simulation rather than a downsampled version of the d02 target, this experiment provides a direct test of emulator behavior on real coarse-resolution input (see Section 4.3).

2.3 Evaluation Regions

We assess emulator performance over six subregions that collectively span the major snow-dominated landscapes of the western United States (Figure S1), following the regional definitions of Rahimi et al. (2022). Four broad regions are considered: the Pacific Northwest (PNW), the Northern Central Rockies (NCR), the Southwest (SW), and the South Central Rockies (SCR). Two smaller regions are also included: the Sierra Nevada (SN), centered on the Lake Tahoe area, and the Northern Rocky Mountains (NRM), encompassing the greater Yellowstone region. These six regions sample the diversity of western US snow regimes: the PNW and SN represent maritime snowpacks characterized by warm temperatures, high precipitation rates (Serreze et al., 1999; Trujillo and Molotch, 2014), and deep seasonal accumulation; the NCR and NRM represent continental snowpacks where cold conditions favor snow persistence but lower precipitation limits accumulation, and where snowpack storage remains sensitive to both temperature and precipitation variability (Luce et al., 2014); and the SCR and SW represent intermountain and arid-mountain environments where snow accumulation is more episodic and marginal (Petersky and Harpold, 2018). Meanwhile, the SN and NRM subregions contain some of the deepest seasonal snowpacks in the domain and are critical to regional water resource management (Pederson et al., 2011; Yang et al., 2023). Figure S1 also shows the approximately 90-km coarse input grid points overlaid on the WRF-WUS d02 domain boundary, illustrating the resolution gap between a typical climate model forcing (features) and the 9-km target (labels) grid that the downscaling framework must bridge. Regional SWE metrics are computed by spatially averaging all terrestrial grid cells within each subregion boundary, and all evaluation statistics are reported on a daily temporal resolution unless otherwise noted.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

3 SWEDE Model Description

3.1 Dual-Encoder Swin U-Net

We predict SWE using the foundational U-Net architecture (Ronneberger et al., 2015; Cao et al., 2022), which processes high-resolution static physiographic variables and coarse-resolution dynamic atmospheric fields through separate encoding branches before merging them in a shared decoder (Figure 1). This asymmetric design is motivated by the distinct spatial characteristics of the two input types. Static variables such as topography and land cover are available at the target 9-km resolution and primarily contain fine-scale local structure, whereas dynamic meteorological fields from ESMs are provided at approximately 90-km resolution and reflect broader mesoscale to synoptic-scale patterns. Using dedicated encoders allows each input stream to be represented at its native scale prior to fusion.

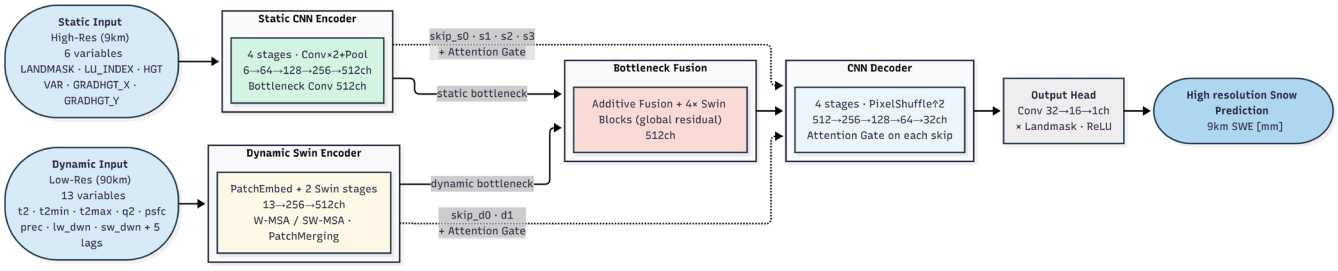


Figure 1. Architecture of the Dual-Encoder Swin U-Net for predicting 9 km SWE from multiscale inputs. High-resolution static land-surface and terrain variables are processed by a convolutional encoder, while low-resolution dynamic atmospheric variables are processed by a Swin Transformer encoder. The two feature streams are fused at the bottleneck and decoded to produce high-resolution SWE predictions.

The static encoder is a four-stage convolutional network. Each stage contains two 3 by 3 convolutional layers followed by Group Normalization (Wu and He, 2018) and Sigmoid Linear Unit (SiLU) activation (Elfwing et al., 2018; Ramachandran et al., 2017), with 2 by 2 max pooling used for downsampling. The number of channels increases from 64 to 128, 256, and 512 across the four stages, and the intermediate feature maps are retained as skip connections for the decoder. The dynamic encoder is a two-stage Swin Transformer (Liu et al., 2021). It begins with convolutional patch embedding and then applies successive Swin Transformer stages with Patch Merging downsampling. Window-based self-attention is used within each stage, with shifted-window operations (Liu et al., 2021) in alternating layers to enable information exchange across neighboring windows while retaining computational efficiency.

At the bottleneck, the deepest features from the two encoders are projected to a common channel dimension and fused by element-wise addition. The fused representation is then refined by residual Swin Transformer blocks with a global skip connection before entering the decoder. The decoder progressively restores the fused features to 9 km resolution through four upsampling stages based on PixelShuffle (Shi et al., 2016). At each stage, skip features from both encoders are combined with

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

the decoder state, and attention gates (Oktay et al., 2018) are used to adaptively weight the skip information before convolutional refinement.

The output head applies to a final convolutional projection followed by a Rectified Linear Unit (ReLU) activation to enforce non-negative SWE predictions. An optional land mask is further applied to set ocean grid cells to zero. Complete architectural specifications, including patch size, embedding dimensions, number of blocks, attention heads, normalization settings, and regularization choices, are provided in Appendix A.

3.2 Input Variables

The model ingests two groups of input variables corresponding to the two encoder branches. The static branch receives six terrain descriptors on the 9 km target grid (Table S1), chosen to represent the dominant physiographic controls on snow accumulation and persistence without directly encoding geographic location. Latitude and longitude are excluded to reduce the risk of learning location-specific climatological biases and to encourage generalization based on physically meaningful terrain information.

The dynamic branch receives thirteen time-varying daily atmospheric variables at approximately 90 km resolution (Table S1). Eight instantaneous variables describe the daily thermodynamic state, moisture availability, precipitation forcing, and radiative energy balance. Five lagged temporal means represent antecedent atmospheric conditions that influence snowpack evolution over weeks to months. Snowpack on a given day integrates the full season of accumulation and ablation, so instantaneous predictors cannot distinguish identical daily conditions preceded by a wet, cold autumn from those preceded by a dry, warm one. The lagged predictors supply this memory. The five lag windows, spanning 45 to 130 days, were selected by maximizing the mutual information between each candidate window mean and peak SWE, then rounded to the nearest 5 days.

3.3 Loss Function Design

The regional distribution of SWE is extremely zero-inflated. In the ERA5-driven simulation spanning 1950 to 2024, approximately 83 percent of pixels record zero SWE and 86 percent fall below 1 mm on any given day, while the right tail extends beyond 4000 mm in high-elevation accumulation zones. Predicting SWE requires simultaneously solving two coupled problems: a detection problem (where is snow present?) and a regression problem (how much snow is there?). These tasks must be addressed under severe class imbalance, a target range spanning four orders of magnitude, and physical constraints on where snow can plausibly exist. We construct a composite loss with four components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{core}} + \lambda_{\log} \cdot \mathcal{L}_{\log 1p} + \sum_i \lambda_{\text{pin},i} \cdot \mathcal{L}_{\text{pinball},i}(\tau_i) + \lambda_{\text{temperature}} \cdot \mathcal{L}_{\text{temperature}} \ ,$$

The core loss ($\mathcal{L}_{\text{core}}$) controls for the dominance of snow-free pixels, which would otherwise make predicting zero everywhere nearly optimal. It combines binary cross-entropy for snow detection with a Smooth L1 regression term computed only on snow-covered pixels. The log-space loss ($\mathcal{L}_{\log 1p}$) addresses the neglect of shallow snowpacks by penalizing relative

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

rather than absolute error, so that ephemeral and marginal snow receives learning effort comparable to deep snow. The quantile loss ($\mathcal{L}_{\text{pinball,i}}$) addresses the systematic underestimation of deep snow that symmetric losses produce on right-skewed distributions, applying asymmetric penalties that weight underestimation more heavily.

The temperature constraint loss ($\mathcal{L}_{\text{temperature}}$) controls for physically impossible snow under warm conditions. It penalizes nonzero SWE in grid cells whose daily minimum temperatures exceed 20 degrees Celsius, threshold derived from the historical WRF-ERA5 simulation, in which no snow-covered cell satisfies this condition. This term does not improve historical accuracy; it enforces physical consistency when the model is driven by ESM forcing well outside the training distribution, where it suppresses spurious snow along the southern domain boundary. The composite design emerged iteratively, with each term added to remedy a failure mode observed in earlier configurations; full formulations, weighting, class balancing, and numerical details are given in Appendix A.

3.4 Model Training

The model is trained using NAdam with decoupled weight decay (Dozat, 2016; Loshchilov and Hutter, 2017) at a base learning rate of 1×10^{-4}. The learning rate follows a cosine annealing schedule with $T_{max} = 100$, preceded by a 5-epoch warmup period starts from 1×10^{-5} (Loshchilov and Hutter, 2016). Figure S2 shows the training and validation curves over 100 epochs. Both the composite loss and the individual monitoring metrics (MAE and RMSE) decrease rapidly during the first 20 epochs and converge to stable values thereafter, with training and validation curves remaining closely aligned throughout, indicating that the regularization strategy effectively controls overfitting despite the large model capacity. The final validation MAE and RMSE are approximately 1 mm and 5 mm respectively. The learning rate schedule ramps up during the warmup phase, holds near the target rate through mid-training, and decays smoothly under cosine annealing in the later epochs. Once trained, SWEDE inference for one simulation year over the full domain requires approximately 2 seconds on a single A100 GPU, compared to approximately 7,500 CPU-hours per simulation year for the full three-stage WRF dynamical downscaling pipeline used to generate the WUS-D3 archive (Appendix A6).

4. Results

4.1 Downscaled SWE Evaluation

First, we test the emulation hypothesis by evaluating SWEDE against WRF-ERA5 for the 2014-2024 held-out evaluation period. Across the six study areas, SWEDE accurately reproduces interannual variability of mean regional SWE. The Pearson correlation (R) for the annual mean SWE time series reaches 0.89 for PNW, 0.92 for NCR, 0.97 for SCR, 0.99 for SW, 0.98 for SN, and 0.94 for NRM, with corresponding mean absolute error (MAE) values of 4.55, 5.16, 1.86, 2.63, 12.19, and 5.72 mm (Figure 2; Table 1). Because April 1 SWE is the benchmark date for operational water supply forecasting in the western United States, we also evaluate the April 1 regional-mean SWE directly (Figure S3). Correlations range from 0.86

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

(PNW) to 0.97 (SW) with MAE values of 7.5 to 32.0 mm, moderately weaker than for the annual mean because the single-date value carries no temporal averaging. Despite the strong correlations, the predictions exhibit a systematic low-variability bias, underestimating high-SWE years and overestimating low-SWE years, most visibly in PNW and NCR. The sign and timing of these anomalies are nonetheless reproduced in all subregions. The model thus correctly identifies which years are anomalously high or low in snow, indicating that it responds to the full range of large-scale forcing conditions encountered during the evaluation period, while compressing the amplitude of that response. The regional correlations of yearly average snow water volume, expressed as integrated totals in km³, range from 0.89 to 0.99 across the six subregions (Figure S4). Because spatial integration removes any influence of zero-inflated grid cells, these correlations directly measure the model's skill in reproducing total snow water amounts rather than snow presence. The low-variability bias identified at the regional-mean scale persists in the integrated volumes: the scatter about the 1:1 line shows that the driest years are overpredicted, most notably in PNW and NCR, while the wettest years are underpredicted, as in SW and SCR (Figure S4). Skill is therefore lowest at both tails of the interannual distribution, consistent with loss-driven regression toward the climatological mean rather than a systematic dependence on warm versus cold conditions. Next, we assess peak snowpack magnitude, since the annual peak approximates the season's maximum water storage and is therefore the quantity most relevant to reservoir operations and seasonal supply forecasts. Evaluated as a percentage bias in daily spatial-maximum SWE averaged annually (Figure 3), the model remains below 10% for all subregions and below 12% for all states, indicating that peak snowpack accumulation is preserved across the domain.

Both the WRF target and SWEDE also track observed interannual variability. Against the independent University of Arizona 4 km SWE reanalysis (Broxton et al., 2019), yearly mean SWE volume correlations reach 0.82 to 0.99 for WRF-ERA5 and 0.78 to 0.94 for SWEDE across the subregions, with SWEDE slightly lower because it is trained to match WRF rather than observations (Figure S5). This agreement with a product generated independently of WRF indicates that the learned mappings capture physical structure rather than target-specific patterns.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

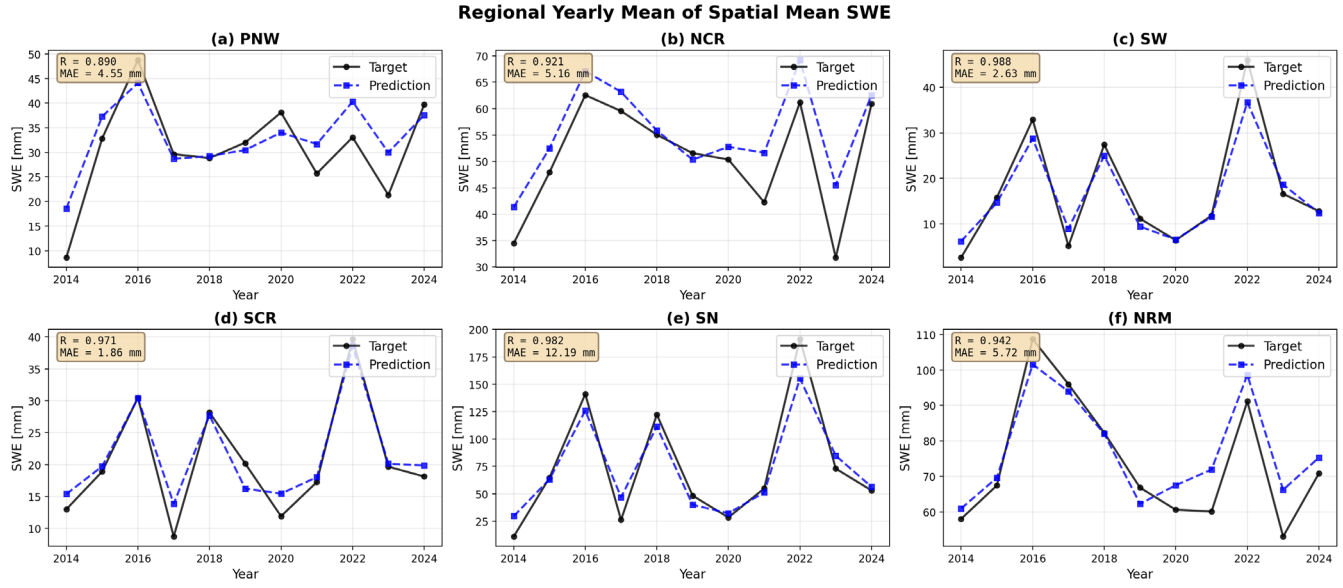
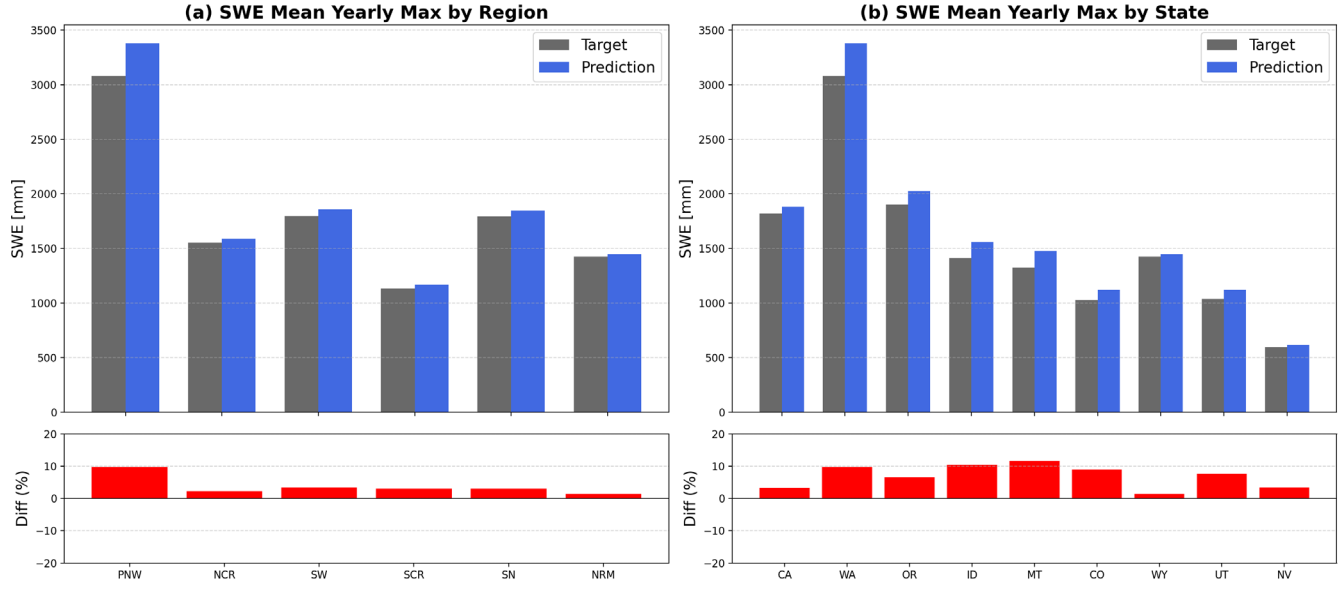


Figure 2. Regional water yearly average of spatial mean SWE (mm) for the six evaluation subregions during the WRF-ERA5 held-out period (2014-2024). Black lines denote the WRF target and blue dashed lines denote the SWEDE prediction. Pearson correlation (R) and mean absolute error (MAE, mm) are annotated in each panel.



https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Figure 3. Spatial Maximum SWE (mm) averaged annually and over the 2014-2024 evaluation period, by subregion (left) and by state (right), comparing the WRF target (grey) and SWEDE prediction (blue). Lower panels show the percentage difference between prediction and target.

Table 1. WRF-ERA5 Evaluation Metrics Summary. Consolidates all evaluation statistics for the WRF-ERA5 held-out period (2014-2024) by subregion.

Region	Interannual r	Max SWE bias (%)	Seasonal r	Spatial r	Pixel MAE (mm)	Snow presence F1
PNW	0.890	9.8	0.995	0.873	10.7	0.823
NCR	0.921	2.2	0.998	0.833	13.7	0.879
SCR	0.971	3.1	0.999	0.832	6.2	0.811
SW	0.988	3.4	0.996	0.812	6.4	0.761
SN	0.982	3.0	0.996	0.851	23.9	0.820
NRM	0.942	1.4	0.999	0.873	15.2	0.881

The statistical fidelity of the downscaled fields extends beyond the first two moments. Quantile-quantile comparison of non-zero SWE values against the WRF target shows that the distributions closely track the one-to-one line across the full range of SWE magnitudes, with MAE from 1.2mm to 6.3mm across all subregions (Figure S6). The only systematic departure occurs at the uppermost quantiles, where the model slightly underpredicts the most extreme SWE values in all subregions. This underprediction is consistent with the conditional-mean smoothing discussed above: the deepest snowpacks occupy a sparse tail of the training distribution, so the pixel-wise loss offers little penalty for regressing these rare values toward more common magnitudes. This agreement is particularly notable given the zero-inflated and heavy-tailed character of SWE, which many regression-based approaches tend to compress toward the mean. In addition, the climatological seasonal cycle of regional SWE tracks the WRF target with seasonal cycle correlations of 0.995 or higher across all subregions (Figure 4; Table 1). For all regions SWEDE and the target also agree on the month of peak SWE, which is March for all regions except NRM and SN (April peak).

To verify that the model also captures the behavior of peak snowpack, we evaluate the seasonal cycle of the spatial maximum SWE within each subregion (Figure S7). The spatial-max seasonal cycle is likewise well reproduced, with correlations of 0.995 to 0.999 across all subregions. The spatial maximum peaks one month later than the regional mean in most subregions (April for PNW, NCR, SW, SCR, and SN; May for NRM), and SWEDE matches this timing in every case. The model reproduces both the accumulation phase from October through March and the ablation phase from April through September. One feature of the ablation phase, the persistence of nonzero spatial-maximum SWE into late summer and its abrupt drop between August and the following September, is a known artefact of the training data. The WRF simulations were run as independent, parallelized years, each initialized to a snow-free state on 1 September, so high-elevation snow still present the previous August is reset at the year boundary rather than carried forward. SWEDE is trained to reproduce WRF and therefore inherits this cold-start discontinuity, so the September fields reflect the initialization rather than a physical late-

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

summer snowpack. Taken together, the annual volume and seasonal cycle comparisons indicate that SWEDE captures not only the statistical distribution of SWE but also the temporal dynamics of snow accumulation and ablation that reflect the underlying physical processes.

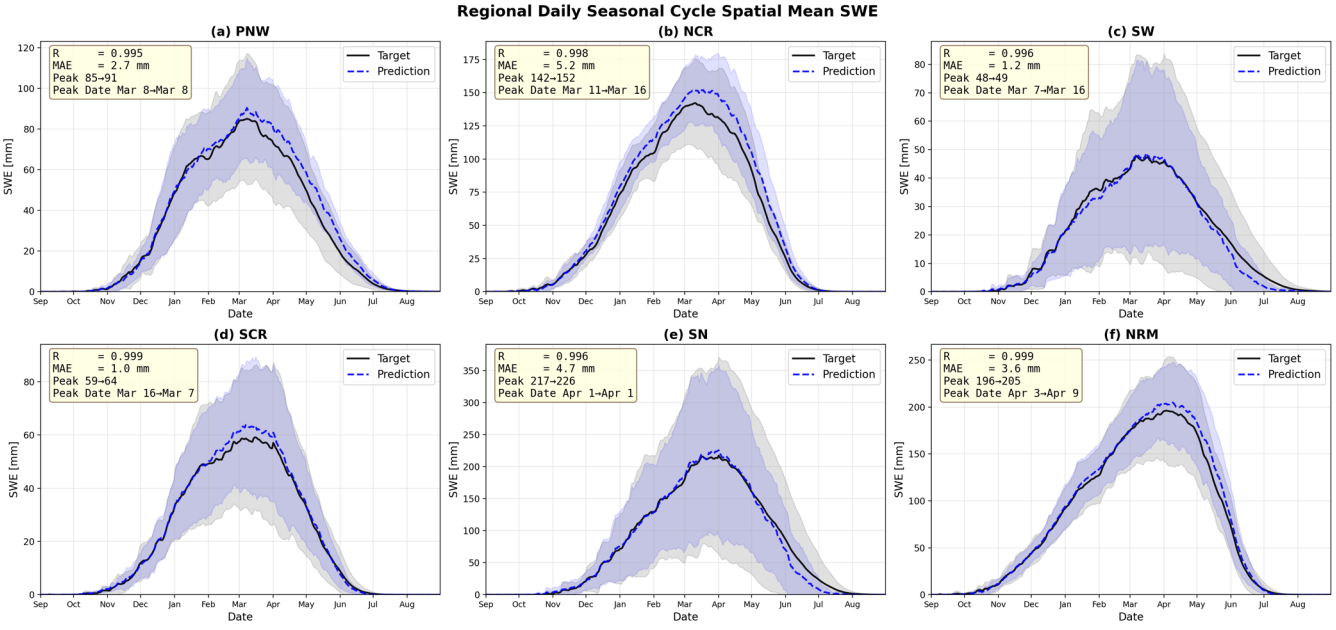


Figure 4. Seasonal cycle of regional mean SWE (mm) from September through August for the six evaluation subregions during the WRF-ERA5 held-out period (2014-2024). Black lines denote the WRF target and blue dashed lines denote the SWEDE prediction. Shading represents the interannual range. Pearson correlation (r), MAE (mm), peak SWE values (mm), and peak date are annotated in each panel.

Spatial fidelity of SWEDE is evaluated through maps of climatological annual maximum SWE and pixel-level scatter analysis. The mapped bias in annual maximum SWE shows that SWEDE reproduces the spatial pattern of deep snowpack across the major mountain ranges of the western United States (Figure 5). The domain-mean bias in climatological annual maximum SWE, averaged over all land pixels, is small (+1.7 mm, +1.8%) and while two-thirds of land pixels fall within ±10 mm, a systematic over-prediction emerges over deep snowpack (climatological annual maximum > 100 mm) areas, averaging +33 mm (+8.5%). This positive bias is largest in the NCR (+44 mm, +12%) and along the Cascade crest (+47 mm, +10%), with comparable values in the NRM (+37 mm, +8%) and SN (+33 mm, +5%). The southern Rockies are likewise biased modestly high (+25 mm, +9%), while the only region with a net negative bias is the arid Southwest (−2.5 mm, −3.4%), where much of the area rarely accumulates snow. The widespread red shading in Figure 5c is compatible with these positive mean biases, because small negative errors in shallow-snow areas can produce large negative percentages, while larger positive

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

errors over deep snowpack can outweigh them in the domain average. The largest relative differences in Figure 5 concentrate along the margins of snow-covered areas. Part of this pattern is arithmetic, since climatological SWE is small there and a modest absolute error yields a large percentage difference, but it also reflects the composition of the training sample: pixels that alternate between snow-covered and snow-free states are rare compared to those that remain snow-free year-round or hold their snow for months at a time, so even a training record spanning more than six decades provides relatively few examples of the transitional regime the model must learn in these zones. To the extent this error stems from sample scarcity rather than a structural limitation of the model, a larger training set, drawn from additional years or additional simulations, may reduce it.

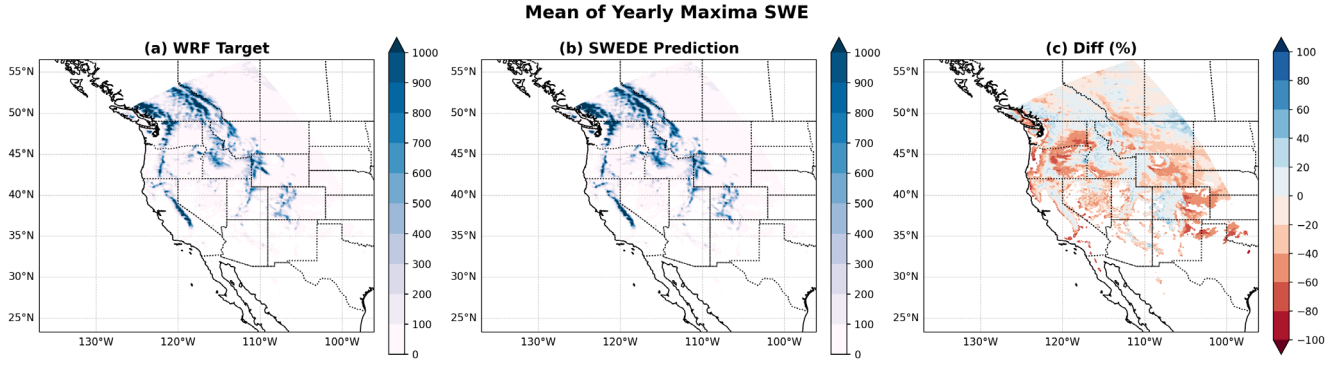


Figure 5. Mean of yearly maximum SWE (mm) for the WRF-ERA5 held-out period (2014-2024): (a) WRF-ERA5 target, (b) SWEDE prediction, and (c) percentage difference between prediction and target. Blue shading indicates overestimation and red shading indicates underestimation.

At the pixel level, the Pearson correlation between SWEDE and target temporal mean over evaluation period SWE across all grid cells is 0.996 with an MAE of 1.98 mm and RMSE of 8.36 mm, and the corresponding correlation for mean yearly maximum SWE is 0.996 with an MAE of 9.55 mm and RMSE of 25.14 mm (Figure S8). Because these statistics are computed over all grid cells, including the large fractions that remain snow-free, they partly reflect the ease of matching zero against zero. A stricter measure evaluates the daily fields directly, without the temporal averaging used above, and only on snow-covered samples (target SWE > 0.1 mm). On this basis the pixel-level MAE is 32.6 mm and the RMSE is 70.3 mm, with the MAE approximately 26 percent of the mean snow-covered value of 123.5 mm. Although sizable relative to the mean, this residual is comparable to the uncertainty inherent in observational SWE itself: gridded SWE products, reanalysis, and model simulations disagree substantially over the mountainous western United States, with both SWE amounts and historical trends depending strongly on model resolution and product choice (Dixit et al., 2026), confirming that the model retains meaningful magnitude skill where snow is actually present and not only on the zero background.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

These high spatial correlations confirm that SWEDE faithfully reproduces the temporal-mean SWE distribution across the western US. SWEDE also shows strong skill in capturing the spatial pattern of SWE within each subregion, with spatial correlation coefficients ranging from 0.81 in the Southwest and 0.83 in the South Central Rockies to 0.87 in both the Pacific Northwest and the Northern Rockies, with the Sierra Nevada at 0.85 and the North Central Rockies at 0.83 (Table 1). Pixel-level MAE is highest in the SN (23.9 mm) and NRM (15.2 mm), reflecting the large absolute SWE values in these deep-snowpack areas, and lowest in the SCR (6.2 mm) and SW (6.4 mm). The mean bias stays small in every subregion, at most about 10% of the regional mean SWE, positive in the PNW, NCR, SCR, and NRM (+4% to +10%) and negative in the SN (−2%) and SW (−5%). The bias is therefore small relative to regional SWE magnitudes. These per-pixel temporal correlations are expected for a daily prediction of a zero-inflated variable at 9 km resolution where local stochastic variability is high, and the strong regional, seasonal, and spatial-pattern metrics confirm that the model captures the aggregated signals relevant to water-resource applications. Non-zero SWE (> 0.1 mm) occurs on 14 to 43 percent of land pixel-days across the subregions during the held-out period, against 20 percent over all domain land, so the subregional metrics are computed on samples less dominated by zeros than the domain-wide statistics.

In addition to continuous SWE evaluation, the model's ability to discriminate snow-covered from snow-free conditions is assessed through binary classification metrics. The domain-averaged F1 score is 0.84, with subregional values ranging from 0.76 in the SW to 0.88 in the NRM (Table 1), indicating that the binary cross-entropy term of the composite loss function contributes effectively to snow presence detection. Snow-covered extent warrants separate evaluation because it governs surface albedo and land-atmosphere energy exchange, and because it is the snow property most directly monitored by satellite remote sensing (Hall et al., 2002). Beyond per-pixel classification, aggregate snow extent tracks the WRF target closely in time ($r = 0.98$ for domain snow-covered area, SWE > 1 mm), although mean extent is biased high by 27%. The error is one-sided and confined to trace snow: precision of 0.75 against recall of 0.95 indicates that the model recovers nearly all snow-covered pixels but extends thin snow into marginal zones that WRF leaves bare, consistent with Figure 5c, and because these false-positive pixels average 6.8 mm SWE they contribute only 2.4% of predicted snow volume, with the extent bias falling to 6% at a 5 mm threshold. Domain-integrated snow water volume is accordingly well preserved, averaging 124.1 km^3 against a target of 116.5 km^3 (+6.5%, $r = 0.98$). Taken together, these results confirm the emulation hypothesis: the model reproduces interannual variability, spatial patterns, seasonal phenology, and peak accumulation magnitudes across all subregions from atmospheric boundary conditions alone.

4.2 Impact of Training Data Composition on Future Projections

The training composition hypothesis predicts that SWEDE's performance over future periods will degrade if trained only on historical data and that warm-ESM training will outperform cold-ESM training. To test this hypothesis, we train SWEDE under four configurations and evaluate each against all 15 ESM-driven WRF simulations across three time periods: 1980-2024 (historical), 2025-2050 (near-future), and post-2050 (late-century). The four training configurations are: (1) WRF-ERA5 only, (2) a single high-warming ESM (CanESM5, high equilibrium climate sensitivity), (3) a single low-warming ESM

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

(FGOALS-g3, low climate sensitivity), and (4) a leave-one-out configuration that trains on all ESMs except the evaluation target (All-excl.). In each case, the self-trained ESM is excluded from the evaluation averages to ensure a fair comparison. Table 2 summarizes the domain-averaged results, and Figures 6 and 7 display the full ESM by training configuration matrices of MAE and Pearson r for the whole domain across the three periods. MAE is computed on daily SWE fields over all land grid cells, so it runs lower than the subregional values in Table 1. Regional breakdowns are provided in Figure S9 and S10.

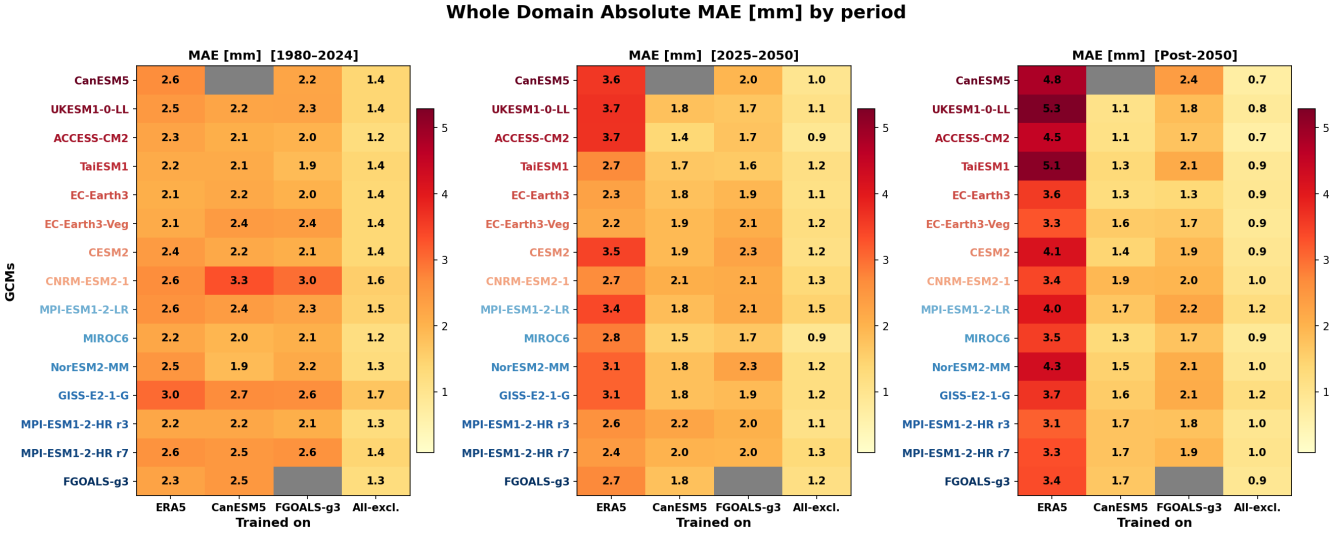


Figure 6. Whole-domain absolute MAE (mm) for SWEDE evaluated on each of the 15 ESM-driven WRF simulations across three time periods: 1980-2024, 2025-2050, and post-2050. Columns correspond to training configurations (WRF-ERA5, CanESM5, FGOALS-g3, and All-excl.), and rows correspond to the evaluation ESM, ordered from warmest to coolest by end-of-century warming over the WUS-D3 domain. Diagonal entries for single-ESM training configurations are excluded. Lower MAE (lighter shading) indicates better performance.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

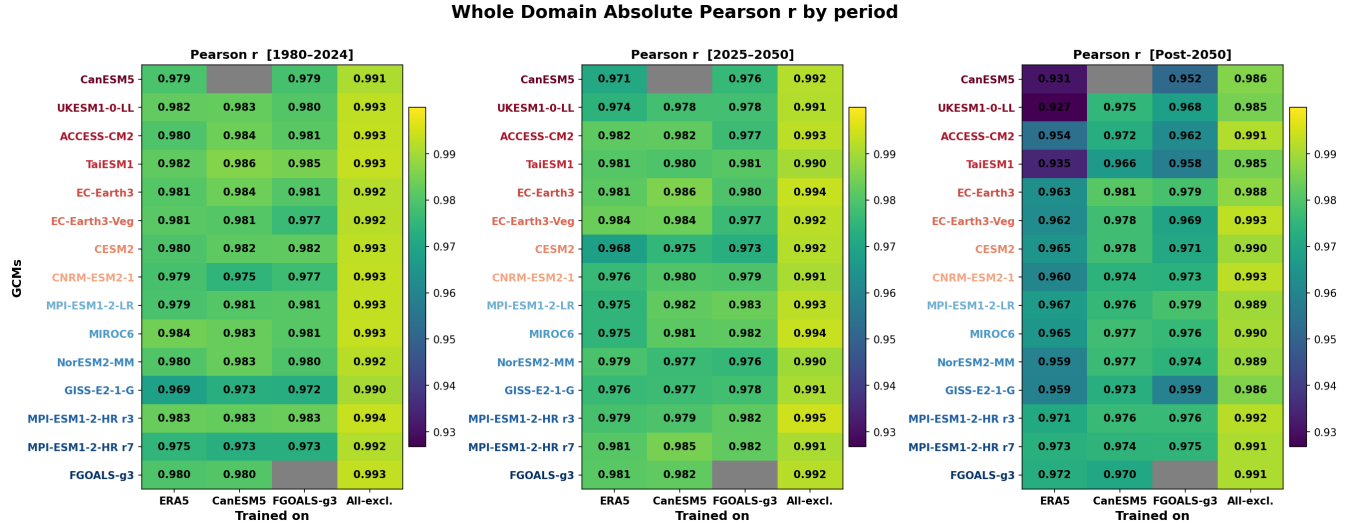


Figure 7. As in Figure 6 but showing Pearson correlation (r). Higher r (lighter shading) indicates better performance.

Table 2. Domain-averaged evaluation metrics for four training configurations, reported as mean values across all ESMs with the self-trained ESM excluded. MAE is in mm and r is the Pearson correlation, both computed on daily SWE fields.

Training configuration	1980-2024 MAE (mm)	1980-2024 (r)	2025-2050 MAE (mm)	2025-2050 (r)	Post-2050 MAE (mm)	Post-2050 (r)
WRF-ERA5 only	2.42	0.980	2.96	0.978	3.96	0.958
CanESM5	2.33	0.981	1.82	0.981	1.50	0.975
FGOALS-g3	2.26	0.979	1.96	0.979	1.91	0.969
All models excl. target	1.38	0.993	1.16	0.992	0.93	0.989

During the historical period (1980-2024), all three single-source training configurations perform comparably across the western US, with MAE values of 2.42 mm (WRF-ERA5), 2.33 mm (CanESM5), and 2.26 mm (FGOALS-g3) and Pearson correlations of approximately 0.98 in all cases (Table 2; Figures S9 and S10). This similarity indicates that when the evaluation climate remains within the range of conditions experienced during training, the choice of training source has a limited effect on prediction skill, as the historical segments of all ESMs and the WRF-ERA5 reanalysis occupy broadly overlapping regions of the input space. Notably, the two ESM-trained configurations draw on about 120 years of simulation (1980-2100) that sample a broader and progressively warming climate, yet their historical skill is indistinguishable from that of the about 64-year WRF-ERA5 record, indicating that record length and climate-regime coverage alone contribute little to in-sample prediction. In contrast, the leave-one-out multi-ESM configuration (All-excl.) reduces historical MAE by roughly 40 percent relative to every single-source training, to 1.38 mm with r of 0.993. The improvement arises from exposure to multiple realizations of internal variability, large-scale circulation, and driving-model precipitation biases rather than from additional

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

years of any single trajectory. Our four configurations do not resolve where this ensemble benefit saturates, and the number of training ESMs required to realize most of the improvement remains an open question with direct implications for emulator training set design.

The training configurations diverge sharply in the post-2050 period, when the climate departs substantially from historical conditions under SSP3-7.0 forcing. The All-excl. configuration remains the best performer by a considerable margin (MAE of 0.93 mm), confirming that multi-ESM diversity provides the most reliable generalization. The WRF-ERA5-only model exhibits the largest degradation, with western US MAE increasing to 3.96 mm and r declining to 0.958, consistent with a distribution-shift interpretation in which the elevated temperatures and reduced snow fractions of the late 21st century lie outside the training envelope (Figures S9 and S10). Training on a single ESM (which includes future data) substantially reduces this degradation, but the choice between a warm and a cold training ESM also matters. The CanESM5-trained model achieves a post-2050 MAE of 1.50 mm compared to 1.91 mm for FGOALS-g3 (Table 2). This advantage of warm-ESM training is consistent across most subregions and is particularly pronounced in the NRM, where the CanESM5-trained MAE of 10.43 mm is substantially lower than the FGOALS-trained value of 16.53 mm (Figure S9). The result indicates that training on an ESM whose future climate spans a broader range of warm, low-snow conditions provides better coverage of the input states encountered in late-century projections, regardless of the specific target ESM.

Time series of predicted SWE for each training configuration, evaluated on the two single-ESM targets, illustrate these dynamics at the subregional level. When FGOALS-g3 is the evaluation target (Figure S11), all training configurations track the WRF target reasonably well throughout the simulation period, because the modest warming of this ESM keeps the future climate relatively close to the historical training distribution. Both the CanESM5-trained and WRF-ERA5-trained models show only mild degradation post-2050, and the All-excl. model matches the target closely in all subregions and periods. When CanESM5 is the evaluation target (Figure S12), the performance separation becomes more apparent: the WRF-ERA5-trained model diverges from the WRF target after approximately 2040, particularly in SN and PNW, while the FGOALS-trained model performs better but still shows larger residuals than when evaluating the cold ESM. The CanESM5-trained model cannot be evaluated on its own target, but the All-excl. configuration tracks the CanESM5-driven WRF simulation with high fidelity throughout the entire period, further demonstrating the value of ensemble-diverse training. These results confirm the training composition hypothesis. As a practical consequence, when only a single ESM is available for training, a warmer ESM is generally preferable to a cooler one for future projection applications.

4.3 Evaluation of SWEDE applied to a separate ESM ensemble

Next, we ask whether a multi-ESM-trained emulator can reproduce both the forced trend and ensemble spread of a fully withheld large ensemble. Based on the findings of Section 4.2, we train SWEDE on the full set of 15 ESMs together with WRF-ERA5 and evaluate on the 10-member CESM2 Large Ensemble. All ten ensemble members are withheld from training, though the training set includes a separate CESM2 experiment among the 15 ESMs, so the test measures generalization to unseen realizations of internal variability under a partially familiar forced response rather than to a fully independent model.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

The SWEDE predictions closely track the WRF-downscaled CESM-LE yearly mean total snow water across all six subregions from 1980 through 2100 (Figure 8). Pearson correlations between SWEDE and WRF ensemble medians are 0.99 for PNW, NCR, and SCR, 0.98 for SW, 0.97 for SN, and 0.96 for NRM, with corresponding MAE values of 0.600, 0.884, 0.429, 0.339, 0.193, and 0.252 km^3, respectively. The agreement is maintained throughout the full simulation period, including the late 21st century when snowpack has declined substantially from historical levels across all subregions. The SWEDE ensemble range closely envelops the WRF ensemble range, indicating that the model generalizes not only to an unseen ESM but also across the ensemble range sampled by the 10-member ensemble.

Figure 8 also illustrates the well-documented limitation of coarse-resolution ESMs in representing mountain snowpack. The native CESM2 SWE (gray shading) is substantially lower than SWEDE fields across all subregions throughout the entire simulation period. This discrepancy is most pronounced in the narrow mountain ranges of the SN and NRM, where the approximately 1-degree CESM2 grid cannot resolve the elevation gradients and orographic precipitation processes that sustain deep snowpacks (Ikeda et al., 2010; Rasmussen et al., 2011; Lundquist et al., 2019). The consistency between SWEDE and WRF in capturing elevation-dependent SWE that the native ESM misses reinforces the value of learned downscaling for producing physically credible SWE projections across large ensembles and suggests that studies relying on raw ESM output may substantially underestimate current snowpack volumes and the absolute magnitude of future snow losses (Pflug et al., 2025; Rhoades et al., 2018).

To test the emulator on real coarse-resolution input rather than the synthetically coarsened d02 fields used in training, we drive it with the 45 km d01 outer-domain output, regridded to the same approximately 90 km input grid so that the comparison isolates the effect of the parent simulation rather than the input grid spacing. The d01-driven prediction preserves the interannual variability and multidecadal decline of the WRF ensemble in all six subregions but is systematically higher, revealing a positive SWE bias when the model is forced from a coarser parent simulation (ML d01 Downscaled in Figure 8). The bias arises because the model learns its amplitude calibration from fields aggregated from the 9 km terrain-resolving simulation, whereas the d01 fields come from a 45 km simulation that resolves orographic processes more coarsely, so the calibration transfers imperfectly even as variability and trend are preserved.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

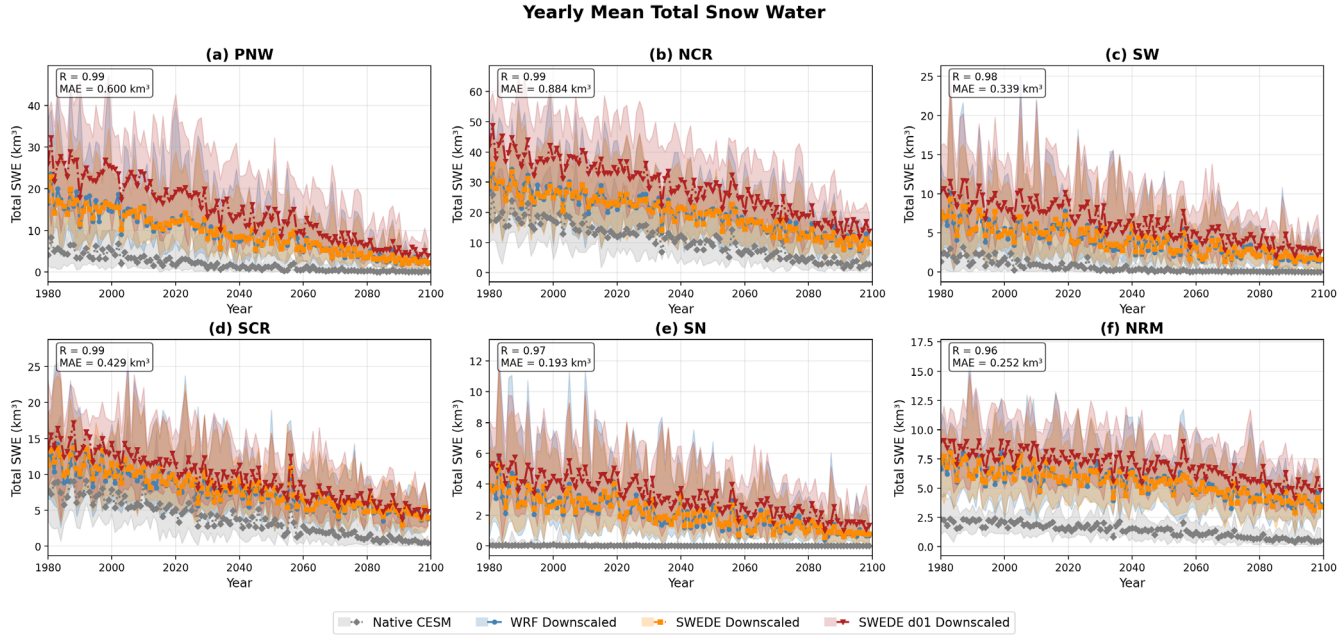


Figure 8. Yearly mean total snow water (km³) for the six evaluation subregions from 1980 to 2100. Blue lines and shading denote the WRF-downscaled CESM2 Large Ensemble median and 10-member range, orange lines and shading denote the SWEDE prediction (ML Downscaled) median and range, red lines and shading denote the SWEDE prediction driven by the 45 km d01 domain (ML d01 Downscaled) median and range, and gray lines and shading denote the native CESM2 CLM median and range. Pearson correlation (r) and mean absolute error (MAE) annotated in each panel are computed between the SWEDE (ML Downscaled) and WRF ensemble medians.

Lastly, we assessed whether SWEDE trained on this collection of models can accurately portray future changes to the seasonal cycle of SWE as simulated by the downscaled CESM2-LE (Figure 9). SWEDE reproduces the WRF-simulated seasonal cycle across all three 30-year climatological windows (1980-2000, 2020-2050, 2070-2100) and all subregions, capturing the progressive decline in peak SWE magnitude and the contraction of the snow season from the historical to the late-century period. By the 2070-2100 window, peak SWE has declined by 24 to 73 percent across the subregions and the melt season terminates earlier, with the largest absolute reductions in NCR and PNW where historical snowpacks hold large amount of volume. The interannual spread across ensemble members narrows from the historical to the late-century window in all subregions. Throughout all three windows, the SWEDE seasonal curves remain closely aligned with the WRF target for both the ensemble mean and the ensemble spread, confirming that the model captures the phenological evolution of snowpack under a changing climate without systematic drift or accumulating bias across the projection horizon. These results confirm the generalization hypothesis: the emulator faithfully captures the progressive decline in peak SWE, the contraction of the snow season, and the narrowing of interannual spread through 2100.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

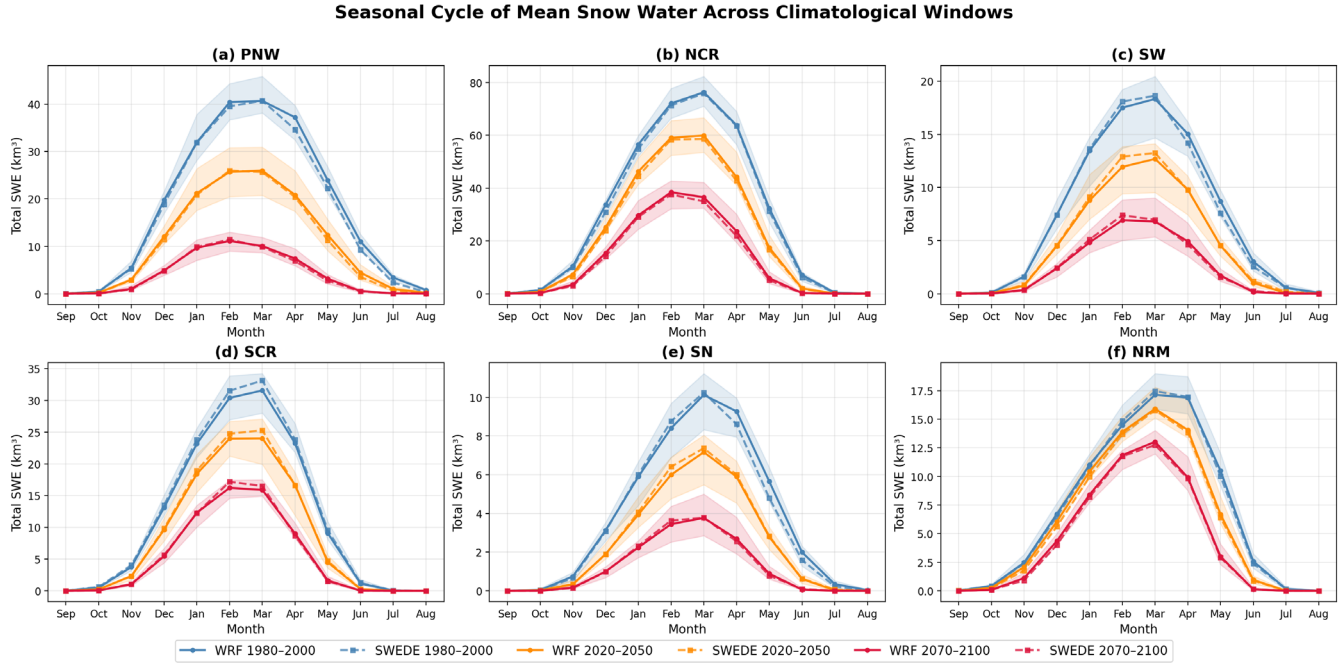


Figure 9. Seasonal cycle of mean total snow water (km³) for the six evaluation subregions, computed over three 30-year climatological windows: 1980-2000 (blue), 2020-2050 (orange), and 2070-2100 (red). Solid lines with circular markers denote WRF-downscaled CESM2 Large Ensemble means, dashed lines with square markers denote SWEDE means, and shading represents the 10-member ensemble spread for each product.

5. Discussion and Conclusions

Here we develop and test a new snow water equivalent downscaling emulator (SWEDE) through a series of training and evaluation experiments spanning historical reanalysis, downscaled CMIP6 projections, and a fully withheld large ensemble. Driven by ERA5, SWEDE agrees closely with both the WRF training target and the independent UA SWE product, establishing that a deep learning model can reconstruct WRF snow processes from atmospheric boundary conditions and terrain alone, without any coarse SWE input. These experiments show that what a model is trained on governs projection robustness more strongly than how it is designed. History-only training degrades systematically under late-century warming, whereas multi-ESM training remains superior across every time period, subregion, and evaluation target. Applied to the withheld CESM2-LE, the multi-ESM model reproduces both the forced trend and the ensemble spread through 2100. This dependence on training data composition is not a peculiarity of our architecture or domain. It reflects a property of supervised learning in non-stationary environments and carries implications for any machine learning downscaling system intended for climate change applications.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

The emulator framing itself merits emphasis, because it distinguishes the present approach from conventional statistical downscaling in a consequential way. SWEDE receives no coarse SWE as input and instead reconstructs snow processes entirely from atmospheric predictors and static terrain descriptors. The strong agreement with both the WRF training target and the independent UA SWE observational product supports the interpretation that the model has learned physically meaningful mappings between atmospheric forcing, terrain, and snowpack response, rather than memorizing a statistical surface anchored to a coarse-resolution version of the target variable. This distinction matters operationally since the model requires only atmospheric boundary conditions. Thus, it can be applied to any ESM that provides standard near-surface meteorological variables, without requiring the driving model to simulate SWE itself.

The physical basis for this training-data sensitivity lies in the threshold nonlinearity of snow-rain partitioning. In marginal elevation zones, modest warming can abruptly shift precipitation from snow to rain, producing snowpack responses that are qualitatively, not merely quantitatively, different from historical conditions (Harpold et al., 2017; Cui et al., 2020). A model trained exclusively on historical data has never encountered these threshold-crossing states and therefore extrapolates poorly into the future. Training on a single high-warming ESM partially addresses this gap by exposing SWEDE to a broader range of warm, low-snow conditions, but training across the full CMIP6 WUS-D3 ensemble proves most effective because it simultaneously samples the diversity of large-scale circulation biases, precipitation climatologies, and climate sensitivities that no individual ESM can span alone. The consistent superiority of multi-ESM training across all time periods, subregions, including the fully withheld CESM2 Large Ensemble, shows that ensemble diversity functions as a form of data augmentation for the non-stationary component of the prediction problem. More broadly, this finding implies that emulators and data-driven forecast models trained solely on historical data degrade under future climates, and that training datasets should instead incorporate counterfactual climates from downscaled projections to remain skillful in a warming world.

A related finding with practical significance is the discrepancy between native ESM and dynamically downscaled SWE. The native CESM2 snow fields substantially underestimate snowpack volumes throughout the simulation period, particularly in narrow mountain ranges where the coarse grid cannot resolve elevation gradients and orographic precipitation (Ikeda et al., 2010; Rasmussen et al., 2011; Lundquist et al., 2019). This underestimation implies that impact assessments relying on raw ESM output may understate both current water storage and the absolute magnitude of future snow losses. A learned emulator such as SWEDE offers a computationally scalable correction by mapping coarse atmospheric boundary conditions to terrain-resolving SWE fields at a fraction of the cost of dynamical downscaling, thereby enabling high-resolution snow projections across large ensembles and multiple scenarios that would otherwise be prohibitively expensive.

Several limitations qualify these findings. SWEDE is trained and evaluated entirely within the WRF modeling framework, so its skill is bounded by the fidelity of the WRF snow physics, and systematic biases in WRF-simulated SWE are inherited by the emulator and propagate into projections. The coarsening procedure used to construct approximate ESM-resolution inputs, 10 by 10 block averaging of WRF fields, does not perfectly replicate the spectral characteristics, spatial coherence, or variable-specific biases of real coarse-resolution model output. The d01 experiment in Section 4.3 quantifies one consequence of this mismatch, because when the model is driven by an independent 45 km simulation rather than a

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.



downsampled d02 field it retains the interannual variability and trend but acquires a positive amplitude bias. Input from a native ESM at roughly 1 degree, produced by a different model with its own biases, is coarser still and remains untested, so a fully operational application would require either retraining directly on native ESM fields where paired high-resolution targets are available or a bias correction calibrated to the target forcing.

Appendix A: Model Implementation Details

A1. Architecture Hyperparameters

Table A1 summarizes the architectural configuration of the SWEDE. The static encoder and decoder each contain four stages, while the dynamic encoder uses two Swin Transformer stages. All convolutional layers in the static encoder and decoder use 3x3 kernels with reflect padding (padding size 1) and no bias when followed by normalization. The base channel count is 64, doubling at each encoder stage and halving at each decoder stage.

Table A1. SWEDE architecture configuration

Component	Parameter	Value
Static encoder	Stages	4
Static encoder	Channels per stage	64, 128, 256, 512
Static encoder	Convolutions per stage	2 (3 × 3, reflect padding)
Static encoder	Downsampling	2 × 2 max pooling
Static encoder	Activation	SiLU
Static encoder	Normalization	GroupNorm
Static encoder	Bottleneck dropout	Dropout2d, 10%
Dynamic encoder	Stages	2
Dynamic encoder	Patch embedding	Convolutional, patch size 4, embedding dimension 256
Dynamic encoder	Swin blocks per stage	2
Dynamic encoder	Attention window size	4 × 4
Dynamic encoder	Shifted window displacement	2 pixels
Dynamic encoder	Attention heads per stage	4, 8
Dynamic encoder	Channel dimensions per stage	256, 512
Dynamic encoder	Multilayer perceptron expansion ratio	4.0
Dynamic encoder	Multilayer perceptron activation	GELU
Dynamic encoder	Multilayer perceptron dropout	10%
Dynamic encoder	Stochastic depth (DropPath)	0 to 0.20, linearly increasing
Dynamic encoder	Downsampling	Patch Merging (2 × 2 concatenation, linear 4C to 2C)
Bottleneck fusion	Projection	1 × 1 convolution, GroupNorm, SiLU
Bottleneck fusion	Fusion method	Element-wise addition
Bottleneck fusion	Residual Swin blocks	4

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Component	Parameter	Value
Bottleneck fusion	Global residual	Yes, around all four blocks
Bottleneck fusion	Attention heads	8
Bottleneck fusion	Window size	4
Bottleneck fusion	Stochastic depth	0.30
Bottleneck fusion	Multilayer perceptron dropout	15%
Decoder	Stages	4
Decoder	Upsampling	PixelShuffle (factor 2)
Decoder	Channel expansion before shuffle	1 × 1 convolution (2 × channels)
Decoder	Post-shuffle refinement	3 × 3 convolution, GroupNorm, SiLU
Decoder	Skip connections	Attention-gated, dual-stream
Decoder	Dynamic skips	Used at the two deepest decoder levels
Decoder	Static skips	Used at all four decoder levels
Output head	Feature refinement	3 × 3 convolution, GroupNorm, SiLU
Output head	Prediction	3 × 3 convolution, SiLU, 1 × 1 convolution (bias = True)
Output head	Land masking	Multiplicative, before activation
Output head	Final activation	ReLU

A2. Attention Gate Implementation

The attention gate at each decoder level computes spatial attention coefficients from the decoder gating signal g and the skip connection input x as follows:

$$g_1 = \mathrm{GroupNorm}\left(\mathrm{Conv}_{1\times1}(g)\right)$$

$$x_1 = \mathrm{GroupNorm}\left(\mathrm{Conv}_{1\times1}(x)\right)$$

$$\Psi = \sigma\left(\mathrm{GroupNorm}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{ReLU}(g_1 + x_1)\right)\right)\right)$$

$$\mathrm{output} = x \cdot \Psi$$

where σ denotes the sigmoid function and all $\mathrm{Conv}_{1\times1}$ layers use no bias. The intermediate projection reduces both g and x to a shared channel dimension equal to half the skip connection channel count. The final 1 × 1 convolution projects to a single channel, producing a spatial attention map that is broadcast across all channels of x. GroupNorm is used instead of BatchNorm within the gates for consistent behavior across batch sizes.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

A3. Normalization Strategy

GroupNorm is used in all CNN components with the number of groups set to $\min(32, C/4)$ where C is the channel count. This ensures a minimum of 4 channels per group, providing stable statistics even at the shallowest encoder stage (64 channels, 16 groups). LayerNorm is used in all Swin Transformer components following standard practice for transformer architectures. BatchNorm is deliberately avoided for two reasons. First, batch statistics are unreliable with geophysical data where batch composition varies seasonally (winter-dominated versus summer-dominated batches have fundamentally different statistical properties). Second, GroupNorm and LayerNorm normalize each sample independently, eliminating train-test discrepancy from batch statistics. The model predicts raw (un-normalized) SWE in millimeters. No output normalization or scaling is applied, so the loss function operates directly in physical units. The ReLU activation enforces SWE ≥ 0 as a hard physical constraint.

A4. Regularization Configuration

Table A2 summarizes the regularization strategy across all model components.

Table A2. Regularization configuration

Technique	Applied to	Setting
Weight decay (NAdam with decoupled weight decay)	Swin patch embedding, encoder, patch merging, and bottleneck fusion weights	0.08
Weight decay (NAdam with decoupled weight decay)	Static CNN encoder and static bottleneck convolutional weights	0.06
Weight decay (NAdam with decoupled weight decay)	Decoder, attention gate, and output head weights (including static skip gates)	0.04
No weight decay	GroupNorm and LayerNorm parameters, all bias terms, and Swin relative position bias tables	0.0
Dropout2d	Static encoder bottleneck	10%
Dropout	Swin encoder multilayer perceptron layers	10%
Dropout	Swin bottleneck fusion multilayer perceptron layers	15%
Stochastic depth (DropPath)	Swin encoder blocks	0 to 20%, linearly increasing
Stochastic depth (DropPath)	Swin bottleneck fusion blocks	20%
Gradient clipping	Global gradient norm	Maximum norm of 1.0
Early stopping	Training termination criterion	Patience of 20 epochs
Land masking	Ocean grid cells in output	Hard constraint

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Weight decay is excluded from normalization parameters and biases following standard practice, as regularizing these parameters interferes with the normalization layers' ability to rescale features (He et al., 2019). The relative position bias tables in the Swin Transformer are also excluded, as they encode geometric priors that should not be shrunk toward zero.

A5. Weight Initialization

Table A3. Weight initialization strategy

Component	Initialization method	Setting
Convolutional layers	Kaiming normal	`fan_in` mode with SiLU-calibrated gain
Convolutional biases	Constant	0
Linear layers in Swin blocks	Truncated normal	Standard deviation of 0.02
Linear biases	Constant	0
GroupNorm and LayerNorm weights	Constant	1
GroupNorm and LayerNorm biases	Constant	0
Relative position bias tables	Truncated normal	Standard deviation of 0.02

Kaiming initialization accounts for the variance scaling of ReLU-family activations, preventing vanishing or exploding gradients in the deep CNN paths (He et al., 2015). Truncated normal initialization for linear layers and position biases follows standard practice for vision transformers (Dosovitskiy et al., 2020).

A6. Computational Environment and Training Cost

All experiments were conducted using PyTorch 2.5.1 with CUDA 12.4 on a single NVIDIA A100 80 GB GPU. The model contains 37.1 million trainable parameters and was trained in single-precision (FP32) arithmetic with a batch size of 60. All training data were preloaded into system memory to eliminate I/O bottlenecks during gradient computation. For the ERA5 baseline configuration (approximately 23,400 daily snapshots spanning 1950-2013), each epoch required approximately 181 seconds, and a full 100-epoch training run completed in approximately 5 hours. For the multi-ESM configurations, which contain substantially more daily snapshots, the combined cost of data preloading and forward-backward computation was considerably larger. These runs were fixed at a 12-hour training budget, during which the model completed approximately 8 epochs. Despite the reduced epoch count relative to the ERA5 configuration, validation metrics stabilized within the allotted training window. Inference for a single daily snapshot over the full 340 × 270 domain requires 3.88 ms, corresponding to approximately 2 GPU-seconds per simulation year. For comparison, the full WRF dynamical downscaling pipeline used to generate WUS-D3 comprises three stages: a 45 km outer-domain simulation (4.5 wall-clock hours on four 36-core nodes), lateral boundary interpolation via ndown.exe (2.5 hours on four 36-core nodes), and a 9 km inner-domain simulation (30 hours on six 36-core nodes), totaling approximately 900,000 CPU-hours and 37 wall-clock hours per 120-year simulation, or roughly 7,500 CPU-hours per simulation year. SWEDE inference for the same simulation year requires approximately 2 seconds on a

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

single GPU. All timing figures reported above reflect computation time exclusive of data loading and preprocessing. For SWEDE, data preloading into system memory prior to training constitutes a non-trivial fraction of total wall-clock time, particularly for multi-ESM configurations where the full training set exceeds 800,000 daily snapshots. Similarly, the WRF pipeline incurs substantial I/O overhead for reading boundary conditions and writing high-frequency output. A fair end-to-end comparison of total wall-clock time would depend on storage architecture, file system bandwidth, and data format choices that vary across computing environments, and is therefore beyond the scope of this study. A fixed random seed of 42 was used for all experiments to ensure reproducibility.

A7. Loss Function Implementation Details

This section provides the complete mathematical specification and numerical stability measures for each loss component described in Section 3.3.

A7.1 Core Loss

The core loss is defined as:

$$\mathcal{L}_{core} = 0.5 \cdot \mathcal{L}_{regression} + 0.5 \cdot \mathcal{L}_{BCE}$$

The binary classification component reinterprets the regression output as an implicit logit for snow presence:

$$\text{logit} = \frac{\log\big((\hat{y} + \varepsilon)/(\tau + \varepsilon)\big)}{T}$$

where $\tau = 0.1$ mm is the snow-presence threshold, $\varepsilon = 10^{-10}$ prevents division by zero, and $T = 1.0$ is a temperature parameter. The logit is clamped to $[-20,20]$ for numerical stability and passed to binary cross-entropy with logits against the binary target $(y > \tau)$.

The dynamic positive-class weight is computed per mini-batch as:

$$w_{pos} = \text{clamp}\left(\frac{n_{\text{no-snow}}}{n_{\text{snow}} + \varepsilon},\ 0,\ 15\right)$$

The upper clamp of 15 prevents extreme weighting in summer batches where fewer than 1% of pixels contain snow.

The regression component uses Smooth L1 loss on snow-covered pixels only:

$$\mathcal{L}_{\text{regression}} = \frac{\sum_{i \in \mathcal{S}} \text{SmoothL1}(\hat{y}_i, y_i;\ \beta = 5.0)}{|\mathcal{S}| + 10^{-8}}$$

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

where $\mathcal{S} = \{i: y_i > \tau\}$. The Smooth L1 transition parameter $\beta = 5.0$ mm means the loss behaves quadratically for errors below 5 mm and linearly for larger errors.

A7.2 Log-Space Loss

$$\mathcal{L}_{\text{log1p}} = \text{mean}\left[\left(\log\left(1+\widehat{y^+}\right) - \log(1+y^+)\right)^2\right]$$

where $\widehat{y^+} = \max(\hat{y}, 0)$ and $y^+ = \max(y, 0)$ ensure non-negative arguments to the logarithm. The weight is $\lambda_{\log} = 0.50$.

A7.3 Multi-Quantile Pinball Loss

For each quantile level τ, the pinball loss on snow-covered pixels is:

$$\mathcal{L}_{\text{pinball}}(\tau) = \frac{1}{|\mathcal{S}| + 10^{-8}} \sum_{i \in \mathcal{S}} (\tau \max(y_i - \hat{y}_i, 0) + (1-\tau) \max(\hat{y}_i - y_i, 0))$$

The three quantile levels (τ) of 0.95, 0.90, and 0.10 are assigned corresponding asymmetry weights (λ) of 0.005, 0.030, and 0.010, respectively.

A7.4 Temperature Constraint Loss

$$\mathcal{L}_{\text{temperature}} = \frac{\sum_{i \in \mathcal{W}} \widehat{y_i^2}}{|\mathcal{W}| + 10^{-8}}$$

where $\mathcal{W}$ is the set of warm-flagged pixels. The warm mask is constructed using the t2min > 20°C as the condition.

The low-resolution temperature fields are bilinearly upsampled to the 9-km prediction grid before applying the mask. The weight is $\lambda_{\text{temp}} = 5.0$.

Appendix B: ESMs List

Table B1. Primary references for the CMIP6 general circulation models and ensemble members used in this study.

Model name	Variant label	Primary citation
ACCESS-CM2	r5i1p1f1	Bi et al. (2020)
CanESM5	r1i1p2f1	Swart et al. (2019)
CESM2	r11i1p1f1	Danabasoglu et al. (2020)
CNRM-ESM2-1	r1i1p1f2	Séférian et al. (2019)
EC-Earth3	r1i1p1f1	Döscher et al. (2021)

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Model name	Variant label	Primary citation
EC-Earth3-Veg	r1i1p1f1	Döscher et al. (2021)
FGOALS-g3	r1i1p1f1	Li et al. (2020)
GISS-E2-1-G	r1i1p1f2	Kelley et al. (2020)
MIROC6	r1i1p1f1	Tatebe et al. (2019)
MPI-ESM1-2-HR	r3i1p1f1, r7i1p1f1	Müller et al. (2018)
MPI-ESM1-2-LR	r7i1p1f1	Mauritsen et al. (2019)
NorESM2-MM	r1i1p1f1	Seland et al. (2020)
TaiESM1	r1i1p1f1	Lee et al. (2020)
UKESM1-0-LL	r2i1p1f2	Sellar et al. (2019)

Code and data availability

The trained SWEDE model code, training scripts, evaluation scripts, plotting scripts and checkpoint are available on Zenodo at https://doi.org/10.5281/zenodo.22326902 (Zhao et al., 2026). All data used for model training and evaluation in this study were obtained from WUS-D3 and are publicly available in the Amazon S3 bucket s3://wrf-cmip6-noversioning/ at https://registry.opendata.aws/wrf-cmip6/ (Rahimi et al., 2024b), accessible through the AWS Open Data Registry. SWEDE is written in Python 3.12 (https://www.python.org/) with PyTorch 2.5.1 (https://pytorch.org/get-started/previous-versions/).

Author contributions

FZ and SR conceived the study and designed the methodology. FZ developed all model and evaluation code, conducted the experiments, analyzed the results, prepared the figures, and wrote the original manuscript. SR supervised the research. CT and JN contributed substantially to manuscript review and editing. MB, KS, AH, and WL also reviewed and edited the manuscript. MB and SR acquired the funding.

Competing interests

At least one of the (co-)authors is a member of the editorial board of Geoscientific Model Development.

Acknowledgements

We would like to acknowledge the use of the Derecho system (doi:10.5065/qx9a-pg09) supported by the NSF National Center for Atmospheric Research (NCAR) at the NSF NCAR-Wyoming Supercomputing Center, sponsored by the National Science Foundation and the State of Wyoming.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Financial support

FZ, MB, and SR acknowledge funding from the U.S. Department of Energy Office of Biological and Environmental Research, Earth and Environmental Systems Sciences Division, through Lawrence Livermore National Lab for the HyperFACETS project," A Framework for Improving Analysis and Modeling of Earth System and Intersectoral Dynamics at Regional Scales" (award no. DE-SC0016605). SR acknowledges the financial support from a National Oceanic & Atmospheric Administration MAPP-NIDIS grant, "Towards predicting drought and subsequent water resource challenges at landscape-resolving scales across the western U.S." (award no. NA230AR4310633). MB and SR also acknowledge the National Science Foundation-supported project, "Wyoming Anticipating the Climate-Water Transition (WY-ACT)" (award no. 2149105). SR, FZ acknowledges support from the NSF ASCEND Engine (award no. 24-0767-A0002-SUB01). CT and JN acknowledge funding from the Regional and Global Model Analysis Program for the Office of Science of the U.S. Department of Energy through the PCMDI project.

References

Bair, E. H., Abreu Calfa, A., Rittger, K., and Dozier, J.: Using machine learning for real-time estimates of snow water equivalent in the watersheds of Afghanistan, The Cryosphere, 12, 1579-1594, 2018.

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

Baño-Medina, J., Manzanas, R., and Gutiérrez, J. M.: Configuration and intercomparison of deep learning neural models for statistical downscaling, Geoscientific Model Development, 13, 2109-2124, 2020.

Barnett, T. P., Adam, J. C., and Lettenmaier, D. P.: Potential impacts of a warming climate on water availability in snow-dominated regions, Nature, 438, 303-309, 10.1038/nature04141, 2005.

Benestad, R. E.: Artificial intelligence and downscaling global climate model future projections, arXiv preprint arXiv:2601.00629, 2026.

Bi, D., Dix, M., Marsland, S., O'farrell, S., Sullivan, A., Bodman, R., Law, R., Harman, I., Srbinovsky, J., and Rashid, H. A.: Configuration and spin-up of ACCESS-CM2, the new generation Australian community climate and earth system simulator coupled model, Journal of Southern Hemisphere Earth Systems Science, 70, 225-251, 2020.

Broxton, P., Zeng, X., and Dawson, N.: Daily 4 km Gridded SWE and Snow Depth from Assimilated In-Situ and Modeled Data over the Conterminous US, Version 1, NASA National Snow and Ice Data Center Distributed Active Archive Center [dataset], 10.5067/0GGPB220EX6A, 2019.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation, European conference on computer vision, 205-218,

Clark, M. P., Hendrikx, J., Slater, A. G., Kavetski, D., Anderson, B., Cullen, N. J., Kerr, T., Örn Hreinsson, E., and Woods, R. A.: Representing spatial variability of snow water equivalent in hydrologic and land-surface models: A review, Water Resources Research, 47, 2011.

Danabasoglu, G., Lamarque, J. F., Bacmeister, J., Bailey, D., DuVivier, A., Edwards, J., Emmons, L., Fasullo, J., Garcia, R., and Gettelman, A.: The community earth system model version 2 (CESM2), Journal of Advances in Modeling Earth Systems, 12, e2019MS001916, 2020.

Dixit, A., Rahimi, S., Huang, L., Musselman, K. N., Vano, J. A., Addor, N., and Lehner, F.: High-resolution models better simulate historical snowpack declines in the Upper Colorado River Basin, Environmental Research Letters, 21, 054012, 10.1088/1748-9326/ae4113, 2026.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Dixon, K. W., Lanzante, J. R., Nath, M. J., Hayhoe, K., Stoner, A., Radhakrishnan, A., Balaji, V., and Gaitán, C. F.: Evaluating the stationarity assumption in statistically downscaled climate projections: is past performance an indicator of future results?, Climatic Change, 135, 395-408, 2016.

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arneth, A., Arsouze, T., Bergmann, T., Bernadello, R., Bousetta, S., and Caron, L.-P.: The EC-earth3 Earth system model for the climate model intercomparison project 6, Geoscientific Model Development Discussions, 2021, 1-90, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, 2020.

Dozat, T.: Incorporating nesterov momentum into adam, 2016.

Dozier, J., Bair, E. H., and Davis, R. E.: Estimating the spatial distribution of snow water equivalent in the world's mountains, Wiley Interdisciplinary Reviews: Water, 3, 461-474, 2016.

Elfwing, S., Uchibe, E., and Doya, K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning, Neural networks, 107, 3-11, 2018.

Fang, Y., Liu, Y., and Margulis, S. A.: A western United States snow reanalysis dataset over the Landsat era from water years 1985 to 2021, Scientific Data, 9, 677, 2022.

Gutmann, E., Barstad, I., Clark, M., Arnold, J., and Rasmussen, R.: The intermediate complexity atmospheric research model (ICAR), Journal of Hydrometeorology, 17, 957-973, 2016.

Hall, D. K., Riggs, G. A., Salomonson, V. V., DiGirolamo, N. E., and Bayr, K. J.: MODIS snow-cover products, Remote sensing of Environment, 83, 181-194, 2002.

Harpold, A. A., Kaplan, M. L., Klos, P. Z., Link, T., McNamara, J. P., Rajagopal, S., Schumer, R., and Steele, C. M.: Rain or snow: hydrologic processes, observations, prediction, and research needs, Hydrology and Earth System Sciences, 21, 1-22, 2017.

He, K., Zhang, X., Ren, S., and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, Proceedings of the IEEE international conference on computer vision, 1026-1034,

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M.: Bag of tricks for image classification with convolutional neural networks, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 558-567,

Ikeda, K., Rasmussen, R., Liu, C., Gochis, D., Yates, D., Chen, F., Tewari, M., Barlage, M., Dudhia, J., and Miller, K.: Simulation of seasonal snowfall over Colorado, Atmospheric Research, 97, 462-477, 2010.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S., Danabasoglu, G., and Edwards, J.: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, Bulletin of the American Meteorological Society, 96, 1333-1349, 2015.

Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., Russell, G. L., Ackerman, A. S., Aleinov, I., Bauer, M., and Bleck, R.: GISS-E2. 1: Configurations and climatology, Journal of Advances in Modeling Earth Systems, 12, e2019MS002025, 2020.

King, F., Erler, A. R., Frey, S. K., and Fletcher, C. G.: Application of machine learning techniques for regional bias correction of snow water equivalent estimates in Ontario, Canada, Hydrology and Earth System Sciences, 24, 4887-4902, 2020.

Klos, P. Z., Link, T. E., and Abatzoglou, J. T.: Extent of the rain-snow transition zone in the western US under historic and projected climate, Geophysical Research Letters, 41, 4560-4568, 2014.

Lee, W. L., Wang, Y. C., Shiu, C. J., Tsai, I., Tu, C. Y., Lan, Y. Y., Chen, J. P., Pan, H. L., and Hsu, H. H.: Taiwan Earth System Model Version 1: description and evaluation of mean state, Geosci. Model Dev., 13, 3887-3904, 10.5194/gmd-13-3887-2020, 2020.

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

Li, L., Yu, Y., Tang, Y., Lin, P., Xie, J., Song, M., Dong, L., Zhou, T., Liu, L., and Wang, L.: The flexible global ocean-atmosphere-land system model grid-point version 3 (FGOALS-g3): Description and evaluation, Journal of Advances in Modeling Earth Systems, 12, e2019MS002012, 2020.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows, Proceedings of the IEEE/CVF international conference on computer vision, 10012-10022,

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.



Loshchilov, I. and Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983, 2016.
Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101, 2017.
Luce, C. H., Lopez-Burgos, V., and Holden, Z.: Sensitivity of snowpack storage to precipitation and temperature using spatial and temporal analog models, Water Resources Research, 50, 9447-9462, https://doi.org/10.1002/2013WR014844, 2014.
Lundquist, J., Abel, M. R., Gutmann, E., and Kapnick, S.: Our skill in modeling mountain rain and snow is bypassing the skill of our observational networks, Bulletin of the American Meteorological Society, 100, 2473-2490, 2019.
Mankin, J. S., Viviroli, D., Singh, D., Hoekstra, A. Y., and Diffenbaugh, N. S.: The potential for snow to supply human water demand in the present and future, Environmental research letters, 10, 114016, 2015.
Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., and Esch, M.: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1. 2) and its response to increasing CO2, Journal of Advances in Modeling Earth Systems, 11, 998-1038, 2019.
Mote, P. W., Hamlet, A. F., Clark, M. P., and Lettenmaier, D. P.: Declining mountain snowpack in western North America, Bulletin of the American meteorological Society, 86, 39-50, 2005.
Mote, P. W., Li, S., Lettenmaier, D. P., Xiao, M., and Engel, R.: Dramatic declines in snowpack in the western US, Npj Climate and Atmospheric Science, 1, 2, 2018.
Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., and Haak, H.: A higher-resolution version of the max planck institute earth system model (MPI-ESM1. 2-HR), Journal of Advances in Modeling Earth Systems, 10, 1383-1413, 2018.
Musselman, K. N., Clark, M. P., Liu, C., Ikeda, K., and Rasmussen, R.: Slower snowmelt in a warmer world, Nature Climate Change, 7, 214-219, 2017.
Najafi, H., Lagerwall, G. L., Obeysekera, J., and Liu, J.: Machine Learning in Climate Downscaling: A Critical Review of Methodologies, Persistent Challenges, and Future Trajectories, 2025.
O'Neill, B. C., Tebaldi, C., Van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.-F., and Lowe, J.: The scenario model intercomparison project (ScenarioMIP) for CMIP6, Geoscientific Model Development, 9, 3461-3482, 2016.
Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., and Kainz, B.: Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999, 2018.
Painter, T. H., Berisford, D. F., Boardman, J. W., Bormann, K. J., Deems, J. S., Gehrke, F., Hedrick, A., Joyce, M., Laidlaw, R., and Marks, D.: The Airborne Snow Observatory: Fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo, Remote Sensing of Environment, 184, 139-152, 2016.
Pan, B., Hsu, K., AghaKouchak, A., and Sorooshian, S.: Improving precipitation estimation using convolutional neural network, Water Resources Research, 55, 2301-2321, 2019.
Pederson, G. T., Gray, S. T., Ault, T., Marsh, W., Fagre, D. B., Bunn, A. G., Woodhouse, C. A., and Graumlich, L. J.: Climatic Controls on the Snowmelt Hydrology of the Northern Rocky Mountains, Journal of Climate, 24, 1666-1687, https://doi.org/10.1175/2010JCLI3729.1, 2011.
Petersky, R. and Harpold, A.: Now you see it, now you don't: a case study of ephemeral snowpacks and soil moisture response in the Great Basin, USA, Hydrol. Earth Syst. Sci., 22, 4891-4906, 10.5194/hess-22-4891-2018, 2018.
Pflug, J. M., Kumar, S. V., Livneh, B., Gutmann, E. D., Gangrade, S., and Kao, S.-C.: Comparisons of montane snow water equivalent projections: calculating total snow mass in regions with projection agreement and divergence in the western United States, Journal of Climate, 38, 855-874, 2025.
Pritchard, H. D., Farinotti, D., and Colwell, S.: Measuring Changes in Snowpack SWE Continuously on a Landscape Scale Using Lake Water Pressure, Journal of Hydrometeorology, 22, 795-811, https://doi.org/10.1175/JHM-D-20-0206.1, 2021.
Rahimi, S., Krantz, W., Lin, Y. H., Bass, B., Goldenson, N., Hall, A., Lebo, Z. J., and Norris, J.: Evaluation of a reanalysis-driven configuration of WRF4 over the western United States from 1980 to 2020, Journal of Geophysical Research: Atmospheres, 127, e2021JD035699, 2022.
Rahimi, S., Huang, L., Norris, J., Hall, A., Goldenson, N., Krantz, W., Bass, B., Thackeray, C., Lin, H., and Chen, D.: An overview of the Western United States dynamically downscaled dataset (WUS-D3), Geoscientific Model Development, 17, 2265-2286, 2024a.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Rahimi, S., Huang, L., Norris, J., Hall, A., Goldenson, N., Risser, M., Feldman, D. R., Lebo, Z. J., Dennis, E., and Thackeray, C.: Understanding the Cascade: Removing GCM Biases Improves Dynamically Downscaled Climate Projections, Geophysical Research Letters, 51, e2023GL106264, https://doi.org/10.1029/2023GL106264, 2024b.

Ramachandran, P., Zoph, B., and Le, Q. V.: Searching for activation functions, arXiv preprint arXiv:1710.05941, 2017.

Rampal, N., Hobeichi, S., Gibson, P. B., Baño-Medina, J., Abramowitz, G., Beucler, T., González-Abad, J., Chapman, W., Harder, P., and Gutiérrez, J. M.: Enhancing regional climate downscaling through advances in machine learning, Artificial Intelligence for the Earth Systems, 3, 230066, 2024.

Rasmussen, R., Liu, C., Ikeda, K., Gochis, D., Yates, D., Chen, F., Tewari, M., Barlage, M., Dudhia, J., and Yu, W.: High-resolution coupled climate runoff simulations of seasonal snowfall over Colorado: a process study of current and warmer climate, Journal of Climate, 24, 3015-3048, 2011.

Rasmussen, R. M., Chen, F., Liu, C. H., Ikeda, K., Prein, A., Kim, J., Schneider, T., Dai, A., Gochis, D., Dugger, A., Zhang, Y., Jaye, A., Dudhia, J., He, C., Harrold, M., Xue, L., Chen, S., Newman, A., Dougherty, E., Abolafia-Rosenzweig, R., Lybarger, N. D., Viger, R., Lesmes, D., Skalak, K., Brakebill, J., Cline, D., Dunne, K., Rasmussen, K., and Miguez-Macho, G.: CONUS404: The NCAR–USGS 4-km Long-Term Regional Hydroclimate Reanalysis over the CONUS, Bulletin of the American Meteorological Society, 104, E1382-E1408, https://doi.org/10.1175/BAMS-D-21-0326.1, 2023.

Rhoades, A. M., Ullrich, P. A., and Zarzycki, C. M.: Projecting 21st century snowpack trends in western USA mountains using variable-resolution CESM, Climate Dynamics, 50, 261-288, 2018.

Rodgers, K. B., Lee, S. S., Rosenbloom, N., Timmermann, A., Danabasoglu, G., Deser, C., Edwards, J., Kim, J. E., Simpson, I. R., Stein, K., Stuecker, M. F., Yamaguchi, R., Bódai, T., Chung, E. S., Huang, L., Kim, W. M., Lamarque, J. F., Lombardozzi, D. L., Wieder, W. R., and Yeager, S. G.: Ubiquity of human-induced changes in climate variability, Earth Syst. Dynam., 12, 1393-1411, 10.5194/esd-12-1393-2021, 2021.

Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention, 234-241,

Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., Decharme, B., Delire, C., Berthet, S., and Chevallier, M.: Evaluation of CNRM Earth system model, CNRM-ESM2-1: Role of Earth system processes in present-day and future climate, Journal of Advances in Modeling Earth Systems, 11, 4182-4227, 2019.

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., Debernard, J. B., Gupta, A. K., He, Y. C., Kirkevåg, A., Schwinger, J., Tjiputra, J., Aas, K. S., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Karset, I. H. H., Landgren, O., Liakka, J., Moseid, K. O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iversen, T., and Schulz, M.: Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations, Geosci. Model Dev., 13, 6165-6200, 10.5194/gmd-13-6165-2020, 2020.

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., O'Connor, F. M., Stringer, M., Hill, R., Palmieri, J., Woodward, S., de Mora, L., Kuhlbrodt, T., Rumbold, S. T., Kelley, D. I., Ellis, R., Johnson, C. E., Walton, J., Abraham, N. L., Andrews, M. B., Andrews, T., Archibald, A. T., Berthou, S., Burke, E., Blockley, E., Carslaw, K., Dalvi, M., Edwards, J., Folberth, G. A., Gedney, N., Griffiths, P. T., Harper, A. B., Hendry, M. A., Hewitt, A. J., Johnson, B., Jones, A., Jones, C. D., Keeble, J., Liddicoat, S., Morgenstern, O., Parker, R. J., Predoi, V., Robertson, E., Siahaan, A., Smith, R. S., Swaminathan, R., Woodhouse, M. T., Zeng, G., and Zerroukat, M.: UKESM1: Description and Evaluation of the U.K. Earth System Model, Journal of Advances in Modeling Earth Systems, 11, 4513-4558, https://doi.org/10.1029/2019MS001739, 2019.

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

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, Proceedings of the IEEE conference on computer vision and pattern recognition, 1874-1883,

Siirila-Woodburn, E. R., Rhoades, A. M., Hatchett, B. J., Huning, L. S., Szinai, J., Tague, C., Nico, P. S., Feldman, D. R., Jones, A. D., and Collins, W. D.: A low-to-no snow future and its impacts on water resources in the western United States, Nature Reviews Earth & Environment, 2, 800-819, 2021.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., Wang, W., Powers, J. G., Duda, M. G., and Barker, D. M.: A description of the advanced research WRF version 4, NCAR tech. note ncar/tn-556+ str, 145, 2019.

https://doi.org/10.5194/egusphere-2026-5377
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.



Sturm, M., Taras, B., Liston, G. E., Derksen, C., Jonas, T., and Lea, J.: Estimating snow water equivalent using snow depth data and climate classes, Journal of Hydrometeorology, 11, 1380-1394, 2010.

Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., and Hanna, S.: The Canadian earth system model version 5 (CanESM5. 0.3), Geoscientific Model Development, 12, 4823-4873, 2019.

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., Sudo, K., Sekiguchi, M., Abe, M., and Saito, F.: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6, Geoscientific Model Development, 12, 2727-2765, 2019.

Trujillo, E. and Molotch, N. P.: Snowpack regimes of the Western United States, Water Resources Research, 50, 5611-5623, https://doi.org/10.1002/2013WR014753, 2014.

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., and Ganguly, A. R.: Deepsd: Generating high resolution climate change projections through single image super-resolution, Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining, 1663-1672,

Walton, D. B., Sun, F., Hall, A., and Capps, S.: A hybrid dynamical–statistical downscaling technique. Part I: Development and validation of the technique, Journal of Climate, 28, 4597-4617, 2015.

Wrzesien, M. L., Pavelsky, T. M., Durand, M. T., Dozier, J., and Lundquist, J. D.: Characterizing Biases in Mountain Snow Accumulation From Global Data Sets, Water Resources Research, 55, 9873-9891, https://doi.org/10.1029/2019WR025350, 2019.

Wrzesien, M. L., Durand, M. T., Pavelsky, T. M., Howat, I. M., Margulis, S. A., and Huning, L. S.: Comparison of methods to estimate snow water equivalent at the mountain range scale: A case study of the California Sierra Nevada, Journal of Hydrometeorology, 18, 1101-1119, 2017.

Wrzesien, M. L., Durand, M. T., Pavelsky, T. M., Kapnick, S. B., Zhang, Y., Guo, J., and Shum, C.: A new estimate of North American mountain snow accumulation from regional climate model simulations, Geophysical Research Letters, 45, 1423-1432, 2018.

Wu, Y. and He, K.: Group normalization, Proceedings of the European conference on computer vision (ECCV), 3-19,

Yang, K., Rittger, K., Musselman, K. N., Bair, E. H., Dozier, J., Margulis, S. A., Painter, T. H., and Molotch, N. P.: Intercomparison of snow water equivalent products in the Sierra Nevada California using airborne snow observatory data and ground observations, Frontiers in Earth Science, Volume 11 - 2023, 10.3389/feart.2023.1106621, 2023.

Zakeri, F., Mariethoz, G., and Girotto, M.: High-resolution snow water equivalent estimation: a data-driven method for localized downscaling of climate data, Hydrology and Earth System Sciences, 29, 6935-6958, 2025.

Zhao, F., Rahimi, S., Bukovsky, M., Singh, K., Thackeray, C., Norris, J., Alex, H., and Liu, W.: SWEDE Model, Zenodo [code], 10.5281/zenodo.22326902, 2026.