Evaluation of Noah-MP Snow Simulation across Site Conditions in the Western United States

MANON VON KAENEL[a] AND STEVEN A. MARGULIS[a]

[a] *Department of Civil and Environmental Engineering, University of California, Los Angeles, Los Angeles, California*

(Manuscript received 27 November 2023, in final form 16 May 2024, accepted 8 July 2024)

ABSTRACT: Quantifying spatiotemporal variability in snow water resources is a challenge especially relevant for regions that rely on snowmelt for water supply. Model accuracy is often limited by uncertainties in meteorological forcings and/or suboptimal physics representation. In this study, we evaluate the performance and sensitivity of Noah land surface model with multiparameterization options (Noah-MP) snow simulations from ten model configurations across 199 sites in the western United States. Nine experiments are constrained by observed meteorology to test snow-related physics options, and the 10th experiment tests an alternative source of meteorological forcings. We find that the base case, which aligns with the National Water Model configuration and uses observation-based forcings, overestimates observed accumulated snow water equivalent (SWE) at 90% of stations by a median of 9.6%. The model performs better in the accumulation season at colder, drier sites and in the melt season at wetter, warmer sites. Accumulation metrics are sensitive to model configuration in two experiments, and melt metrics, in six experiments. Alterations to model physics cause changes to median accumulation metrics from −13% to 2.3% with the greatest change due to precipitation partitioning and to melt metrics from −10% to 3% with the greatest change due to surface resistance configuration. The experiment with alternative forcings causes even greater and wider-ranging changes (medians ranging from −29% to 6%). Not all stations share the same best-performing model configuration. At most stations, the base case is outperformed by four alternative physics options which also significantly impact snow simulation. This research provides insights into the performance and sensitivity of snow predictions across site conditions and model configurations.

SIGNIFICANCE STATEMENT: The purpose of this work is to evaluate the performance and sensitivity of a land surface model's simulation of snow across site conditions and in response to different model configurations. This is important because estimating snow distribution is a challenge especially relevant for regions that rely on snowmelt for water supply. While land surface models can provide useful large-scale estimates, they are often limited by uncertainties in forcings and/or suboptimal physics representation. The results, which show varying model behavior across geography, climate, vegetation types, and model configurations, highlight inadequacies in model physics representation, emphasize the need for accurate meteorological forcings, and suggest that customizing model configurations to the unique characteristics of the domain could yield more accurate and useful results.

KEYWORDS: Model evaluation/performance; Snow; Water resources; Land surface model; North America

1. Introduction

Estimating snow water equivalent (SWE) in remote mountainous areas remains one of the most challenging problems in hydrology (Lettenmaier et al. 2015; Dozier et al. 2016). This challenge is particularly important in areas where snowpack plays a significant role in seasonal water supply and regional hydrology, such as the western United States. To make critical management decisions for flood control, hydropower operations, irrigation, and other competing demands in such regions, water managers need accurate assessments of the space–time distribution and availability of water in snowpack (Vicuña et al. 2011; Tanaka et al. 2006; He et al. 2016). Snowpack observations can come from in situ stations or remote sensing platforms; however, both data sources are limited in time and/or space and are prone to substantial errors, especially in topographically complex areas (Dozier et al. 2016).

Land surface models (LSMs) are used to generate spatially distributed estimates of SWE, snow-derived runoff, and other hydrologic variables over large spatial domains and at varying spatiotemporal resolutions by simulating the physics of the water and energy cycles at the land surface. The accuracy of these modeled estimates depends on the reliability of meteorological forcings as well as the fidelity of model physics (Cho et al. 2022). Various studies (e.g., Cho et al. 2022; Kim et al. 2021; Pan et al. 2003; Barlage et al. 2010) have demonstrated that a common weakness in LSMs is that SWE estimates are highly uncertain and often underestimated, which can cascade into uncertainties and errors in SWE-dependent variables like runoff and evapotranspiration. Properly describing and simulating snow processes in a LSM thus helps to accurately predict not only the distribution of snow water resources but also other variables in the land surface water cycle.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-23-0211.s1.

Corresponding author: Manon von Kaenel, mvonkaenel@g.ucla.edu

DOI: 10.1175/JHM-D-23-0211.1

Here, we examine the performance and sensitivity of SWE estimates produced by the Noah LSM with multiparameterization options (Noah-MP). Noah-MP is the hydrologic core and LSM for the operational National Water Model (NWM) and associated climate predictions based on the Weather Research and Forecasting (WRF) regional climate model. Of the four LSMs tested by Cho et al. (2022), Noah-MP generated the most accurate maximum SWE values, but it nonetheless significantly underestimated observations, and several limitations in current physics options were noted. Past studies have also identified biases in the Noah-MP representation of snow depth, timing of snow disappearance, and annual SWE (e.g., Q. Li et al. 2022; J. Li et al. 2022; You et al. 2020; Garousi-Nejad and Tarboton 2022; He et al. 2021). The design of Noah-MP allows for user-defined selection of options for physical processes such as precipitation partitioning, snow albedo, and vegetation–snow interactions. Past studies have explored how the incorporation of different schemes for some or most of these physical processes in Noah-MP affects the simulation of snow depth or SWE at specific sites (You et al. 2020; Zhang et al. 2016; Letcher et al. 2022; Wang et al. 2019) and at a global scale (Q. Li et al. 2022). However, for the purposes of understanding uncertainties in snow simulation in a regional hydrologic model like the NWM, these studies were limited in scope by the number of sites or model resolution.

In this study, we evaluate the performance and sensitivity of Noah-MP snow simulations across different model configurations at sites spanning the western United States (WUS). The WUS contains extensive in situ data that can force and validate the model and represents a diversity of climate and site conditions. For the NWM, the WUS mountains are also where the snow model plays the largest role in runoff/streamflow prediction. Our intent is to provide insights into model behavior relative to user-selected physics options and promote the improvement of snow simulation in regional climate and hydrologic applications. To this end, we aim to answer the following: 1) How well can the National Water Model configuration of Noah-MP reproduce observed SWE at sites across the western United States? 2) How sensitive are the Noah-MP snow simulations to changes in model configurations? 3) How do model errors and sensitivities vary by region, climate, and vegetation conditions?

2. Data

a. Study sites

The study domain is comprised of 199 stations from the Snowpack Telemetry (SNOTEL) network across the WUS that meet the following criteria: 1) Less than 5% of daily precipitation and hourly temperature observations for November–June over water years (WYs) 2007–19 are missing and 2) no daily SWE observations over the study record are missing. Daily precipitation and SWE records were taken from the bias-corrected quality-controlled product published by the Pacific Northwest National Laboratory (PNNL; Yan et al. 2018), and hourly temperature data were downloaded from the National Resources Conservation Service (NRCS) web portal.

We classified the 199 sites into groups based on geography, climate, and vegetation type to evaluate how the model performs across site conditions that are relevant to snow accumulation and/or melt (Fig. 1). We assigned each station to an ecoregion based on the Commission for Environmental Cooperation (CEC) level III terrestrial ecoregions classification (Wiken et al. 2011) (see Text S1 in the online supplemental material). For illustration purposes, certain nearby ecoregions were combined because model behavior was similar. We also developed climate subgroups based on observed temperature and precipitation using the classification scheme outlined in Sun et al. (2022); stations with mean winter (November–March) temperature less than −1°C are classified as "cold," and stations with winter precipitation less than the 25th percentile of all station values are classified as "dry." Last, each station was assigned a vegetation type from the USGS land-use/land-cover (LULC) classification system based on site photos posted on the NRCS portal.

b. Meteorological forcings and observations

Forcings required by the Noah-MP model at an hourly time step are precipitation, temperature, specific humidity, terrain-level pressure, downward longwave and shortwave radiation, and surface wind magnitude. For the base case and model experiments in which physics options were altered (see section 3b), we assigned hourly temperature and daily precipitation inputs from SNOTEL observations for consistency with the SNOTEL SWE used to evaluate model performance. To generate records that are internally consistent, accurate, unbiased, and complete, extensive quality control, bias correction, and gap-filling procedures were applied (see Text S2). So, to the greatest extent possible, model errors in these experiments can be linked to the model physics rather than forcing biases because the forcings are both constrained by observations and consistent with the validation dataset. Daily precipitation observations were divided into even hourly values to match the model time step; this is a common approach in large-scale hydrologic modeling [e.g., as is used by default in the Variable Infiltration Capacity model (VIC; Liang et al. 1994)] due to challenges in obtaining reliable subdaily records. We completed the forcings for these experiments with the Analysis of Record for Calibration (AORC) version 1.1 (Kitzmiller et al. 2018) dataset, because it was developed by NOAA's Office of Water Prediction (OWP) specifically for the purpose of providing gridded meteorological forcings to calibrate the NWM. It has both high spatial (~800 m) and temporal (hourly for water years 1979–2019) resolutions (see Text S3 for more). We downscaled humidity and pressure from AORC values with SNOTEL temperature and elevation following methods outlined in Liston and Elder (2006) and Cosgrove et al. (2003), respectively. Longwave radiation, shortwave radiation, and wind inputs were extracted from the AORC pixel containing the SNOTEL site. Note here that our decisions to combine AORC hourly radiation forcings with SNOTEL temperature and precipitation, and to divide daily precipitation into even

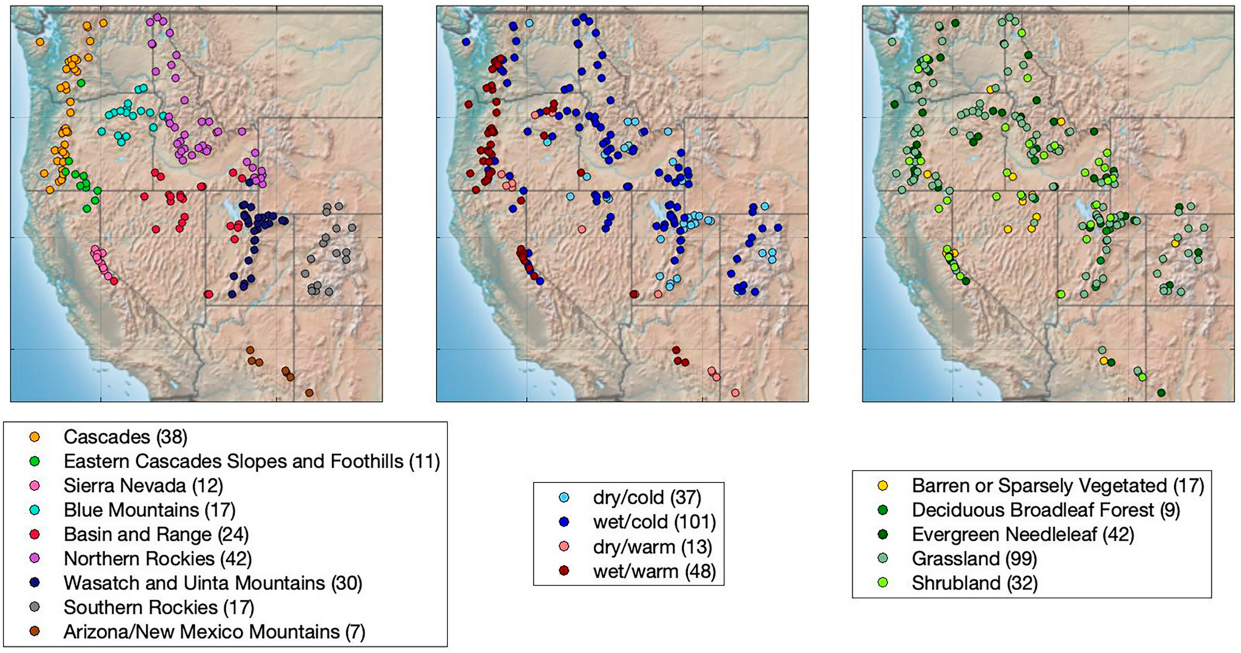

FIG. 1. Map of 199 SNOTEL stations used, categorized by (left to right) geographic region, climate, and vegetation type. The number of stations in each subgroup is noted in the legend entries.

hourly values, will invariably introduce some inconsistencies within the meteorological forcing dataset. However, we believe this was the best approach for the purposes of this study given data limitations and model physics.

To further evaluate model performance and sensitivity relative to changes in meteorological forcings, we set up an additional experiment using solely the gridded AORC dataset, forcing each model point with values from the nearest AORC pixel.

3. Model

a. *Overview of the Noah-MP model*

Noah-MP is a community model that solves the land surface energy and water balances and simulates land–atmosphere fluxes to generate a suite of hydrologic and other land surface variables (Niu et al. 2011). It was developed to improve upon the limitations of the Noah LSM (Gochis et al. 2018). Snowpack is represented by up to three layers. Within each layer, snow density, temperature, and liquid water fraction are computed for each time step, and snow accumulation and ablation processes are based on mass and energy balance. Snow-cover fraction on the ground is a function of snowpack depth and density and tunable parameters (Niu and Yang 2007). Vegetation is represented by a single-layer canopy model with the capability for snow interception and unloading on a canopy snow-cover fraction (Niu et al. 2011). A semitile subgrid scheme evaluates the radiation balance, where shortwave radiation is computed over the entire grid cell using a two-stream approximation and considering canopy gap probabilities, while radiation component fluxes and albedo are computed separately over the vegetated and bare ground areas (Niu et al. 2011). Noah-MP uses multiple user-defined options for key land–atmosphere interaction processes, such that over 4500 total combinations of physics configurations can be assessed. These configurations relate to 14 physics processes (see Table S1) such as dynamic vegetation, precipitation partitioning, soil temperature, and snow albedo (Gochis et al. 2018). This setup allows users to customize their modeling scheme. More details about the model representation of these physics processes can be found in Text S4.

b. *Application of Noah-MP*

We applied Noah-MP as a standalone mode of WRF-Hydro v5.1.1, which is the core physics in NWM v2.1 (Cosgrove et al. 2024), over the 199 SNOTEL sites for the period WYs 2007–19. This period covers average, dry, and wet years. Each site was treated as a single Noah-MP grid cell with a resolution of 1 km to match that of the AORC forcings and NWM v2.1. We manually adjusted the coded vegetation type at each grid cell to match the corresponding SNOTEL station based on site photos from NRCS. When no photos were available, we used the default LULC class from the WRF preprocessing system (WPS) for the corresponding grid cell. All other surface conditions and parameters, such as default snow albedo parameters, soil classification, vegetation leaf area index (LAI), and other static descriptors, were defined by the WPS.

We tested 10 model configurations relevant to snow processes (Table 1), including one base case and nine alternative experiments. The base case uses SNOTEL-based meteorological forcings and has physics options set to match the NWM v2.1 recommended configuration. Seven physics processes were

TABLE 1. List of selected Noah-MP physics processes, with the base case option listed first and alternative option(s) second. The activated set of physics options for the base case and each of the other nine model experiments are also indicated. AORC is the only experiment that shares the base case configuration of physics options. Note that three versions of the snow albedo BATS model were tested: a case with all default parameter values (base case), a case with the snow age parameter $\tau_0 = 3.05 \times 10^6$ ($BATS_{tau\_vis}$), and a case with $\tau_0 = 5.29 \times 10^5$ ($BATS_{tau\_NIR}$) based on optimized values from Abolafia-Rosenzweig et al. (2022). The symbol * shows these model experiments are ones that primarily impact snow accumulation processes; the symbol ** shows these model experiments are ones that primarily impact snowmelt processes; and the symbol *** shows these model experiments are ones expected to impact both snow accumulation and snowmelt processes.

Noah-MP physics process	Options	Model configurations									
		Base case	Precip2.2*	Precip0*	Alb**	TempLB**	TempSolv**	ResisEvap**	ResisDrag**	DynVeg***	AORC***
Precipitation partitioning	Jordan (1991): prescribed linear snowfall fraction when air temperature > 0.5°C and <2.5°C	x			x	x	x	x	x	x	x
	Air temperature < 2.2°C		x								
	Air temperature < 0°C			x							
Snow albedo	BATS	x	x	x		x	x	x	x	x	x
	CLASS				x						
Lower boundary condition of soil temperature	Temperature at soil lower boundary (8 m) read from file	x	x	x	x		x	x	x	x	x
	Zero heat flux from bottom					x					
Snow/soil temperature time scheme	Semi-implicit; flux top boundary condition; fractional snow cover used in calculation	x	x	x	x	x		x	x	x	x
	Semi-implicit; flux top boundary condition						x				
Surface resistance to evaporation/ sublimation	Sakaguchi and Zeng (2009) for nonsnowy pixels; parameter read from file for snowy pixels	x	x	x	x	x	x		x	x	x
	Sakaguchi and Zeng (2009)							x			
Surface layer drag coefficient	Monin–Obukhov similarity theory	x	x	x	x	x	x	x		x	x
	Original Noah [Chen et al. (1997)]								x		
Dynamic vegetation	Module turned off (vegetation parameters read from file)	x	x	x	x	x	x	x	x		x
	Module turned on (prognostic vegetation growth)									x	
Forcings	SNOTEL temperature and precipitation	x	x	x	x	x	x	x	x	x	
	AORC										x

examined because of their relevance to snow simulation. You et al. (2020) found that modeled snow depth was sensitive to six of these processes at most sites. The seventh process was included because it offers a specific adjustment for snowy pixels. The tested physics processes (and named experiments) are precipitation partitioning (Precip2.2 and Precip0), snow albedo (Alb), lower soil temperature boundary condition (TempLB), snow/soil temperature time scheme (TempSolv), surface resistance to evaporation/sublimation (ResisEvap), surface layer drag coefficient (ResisDrag), and dynamic vegetation (DynVeg) (Table 1). Additional details on these physics processes and their alternative options are provided in Text S4. The physics processes tested by the alternative model configurations impact snow accumulation and/or melt processes through various physical mechanisms (Fig. S2). For the physics processes with multiple alternative options, we tested only the recommended Noah-MP alternative (Gochis et al. 2018). Two alternatives were tested for the precipitation partitioning process due to its relevance to snow accumulation. The last experiment (AORC) uses the base case model configuration but AORC forcings. For this experiment, discrepancies between observed SWE and Noah-MP simulations can be attributed to both meteorological input and model errors.

In addition to these user-defined physics options, Noah-MP employs parameters, some of which are designed to be tunable by the model user, related to vegetation and soil classes or model processes like runoff, albedo, and radiation balances. These parameters are not the focus of this study, but it is worth noting that particularly relevant for snow simulation are those parameters related to snow albedo (e.g., Sun et al. 2019; Abolafia-Rosenzweig et al. 2021; He et al. 2021). The Biosphere–Atmosphere Transfer Scheme (BATS), which is the default snow albedo model in Noah-MP and represents both direct and diffuse albedo over visible and near-infrared (NIR) spectra (Dickinson et al. 1986) in a more sophisticated way than the alternative Canadian Land Surface Scheme (CLASS) model, has 12 tunable parameters. Abolafia-Rosenzweig et al. (2022) demonstrate that the BATS model is significantly sensitive to some of these parameters and provide parameter values that are locally optimized to stations in the southern Rockies. Letcher et al. (2022) further find that the default parameterization of the BATS model yields albedo values that underestimate the rate of observed snow aging. So, to acknowledge the potential breadth of performance of the BATS model and the caveats of its default parameters, we further test two alternative values for the empirical snow age parameter τ_0, to which Abolafia-Rosenzweig et al. (2022) find BATS is especially sensitive. The experiment hereafter called $\text{BATS}_{\text{tau\_vis}}$ uses $\tau_0 = 3.05 \times 10^6$ to match the value optimized by Abolafia-Rosenzweig et al. (2022) for visible albedo, and the experiment $\text{BATS}_{\text{tau\_NIR}}$ sets $\tau_0 = 5.29 \times 10^5$, the value optimized for NIR albedo. These two values are, respectively, greater and less than the default value of $\tau_0 = 1 \times 10^6$. Although these values were optimized locally in the southern Rockies, we apply them in a global fashion across all sites to evaluate 1) the sensitivity of the BATS albedo to this parameter and 2) how well these parameter values perform at both similar and dissimilar sites. Note that all other parameter values and physics options in these two experiments otherwise match the base case (Table 1).

4. Methods

We computed snow metrics to assess volume, rates, and timing of snow simulation across both the accumulation and melt seasons. These snow metrics, illustrated in Fig. 2 and described in more detail in section 4a, are used to evaluate both model performance relative to SWE observations and sensitivity of the different model experiments (Table 2).

a. Derived annual SWE metrics

For each station year, we computed four annual snow metrics from daily time series of observed and simulated SWEs as illustrated in Fig. 2: accumulated SWE (mm), storm rate (mm day^{-1}), peak SWE day of water year (DOWY; day), and daily melt rate (mm day^{-1}).

We calculated the accumulated SWE by summing all positive daily changes in SWE (Fig. 2a), to represent the total snow available for melt during the water year. This integrated metric is crucial for operational hydrologic models like the NWM to correctly predict water supply. We chose this metric over the traditional single-day metric of peak SWE because peak SWE underestimates total snow availability, as it misses winter melt events and postpeak accumulation (e.g., Fig. S3). The day of peak SWE (Fig. 2a) is included to quantify the timing of spring melt onset.

We computed the annual-averaged daily melt rate to independently characterize the snowmelt process. The daily melt rate is the average negative change in SWE over a melt window. The melt window spans from the water year's first day to the snow-off day (Fig. 2b). Note that this window excludes late-season accumulation/melt events that happen after the winter snowpack has melted out. The SNOTEL snow pillow's precision is 2.54 mm, so daily SWE changes below the propagated error of 3.58 mm were set to zero in both simulated and observed records to maintain consistency (see more in Text S5). Only events exceeding a second minimum threshold were included to avoid biasing the average with small rates; this minimum threshold is set for each station as the rate above which 90% of all observed melt occurs. When comparing melt rates from two different time series (i.e., simulated vs observed for the model performance analysis or simulated vs simulated for the model sensitivity analysis), the average is applied over a melt window constrained by the earliest snow-off day (Fig. 2b), because high late-season melt rates in one time series (e.g., Fig. 2d) could bias the comparison if the other time series has already melted out. This way, we are comparing melt rates over the period in which the model and measurements agree that snow is present and melting.

The storm rate represents the average rate of snow accumulation on large storm days. Large storm days are defined as days when the daily positive change in SWE exceeds the historical observed 75th percentile. Serreze et al. (2001) found that in the western United States, this top quartile represents

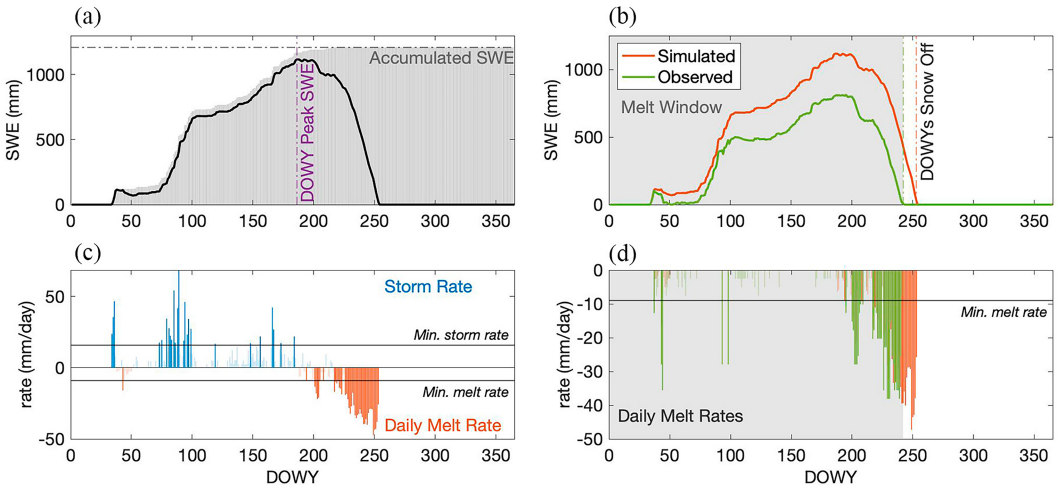

FIG. 2. Illustration of how the four snow metrics are calculated for a sample SWE time series for 1 WY. (a) Peak SWE DOWY and accumulated SWE. (b) Illustration of how a melt window is calculated when comparing two SWE time series for a sample WY. Here, snow-off day is the first day after peak SWE when daily snowmelt is less than 1 mm. The 1-mm threshold discounts late-season small changes in simulated SWE. (c) Storm rate and daily melt rate (averaged over the year), with a sample minimum melt and storm rates indicated. (d) Simulated and observed daily melt rates within the shaded overlapping melt window.

over half of total snowfall. Because large storms contribute significantly to snow accumulation and mountain hydrology in general—for example, atmospheric river events generate 4 times more daily SWE than nonatmospheric river (non-AR) storms (Guan et al. 2010), it is important to evaluate the model's ability to recreate those large snowfall events. Note that because the model is run on an hourly time step but generates daily output, the storm rate could be affected by intraday melt events. For example, melt that occurs on a storm day could reduce the storm rate if snowmelt leaves the snowpack; this could also occur in the observational time series.

b. *Evaluation of model performance*

We assessed model performance across both the accumulation and melt seasons with six performance metrics (Table 2), by comparing simulated to observed SWE over 2388 station years of results. For each year at each site, we computed the bias as a normalized percent between simulated and observed snow metrics M:

$$100\% \times (M_{\text{sim}} - M_{\text{obs}})/M_{\text{obs}}. \tag{1}$$

We then averaged these biases across all years to derive a mean normalized bias (MNB) for each snow metric and station to evaluate the systematic bias in the model. Years for which the bias magnitude was greater than 1000% (during low to no snow years) were removed as outliers before aggregating. To evaluate error magnitude, we derived a mean absolute error (MAE), expressed as a normalized percentage, by averaging yearly absolute errors for each site and snow metric:

$$100\% \times \frac{|M_{\text{sim}} - M_{\text{obs}}|}{M_{\text{obs}}}. \tag{2}$$

TABLE 2. Summary of snow, performance, and sensitivity metrics. Annual snow metrics are computed for both observed and modeled results at each station. Performance metrics characterize systematic bias and error magnitude by comparing modeled (base case) to observed snow metrics at each station. Sensitivity metrics characterize bias relative to the base case and sensitivity level at each station and for each experiment. The symbol * shows these metrics relate to the accumulation season; the symbol ** shows these metrics relate to the melt season.

	Performance metrics		Sensitivity metrics	
Annual snow metrics	Systematic bias	Error magnitude	Systematic bias	Sensitivity
Accumulated SWE (mm)*	MNB (%)	MAE (%)	MNB with respect to the base case (%)	KS statistic
Storm rate (mm day^{-1})*				
Daily melt rate (mm day^{-1})**				
Peak SWE DOWY (day)*	Difference (day)	Absolute difference (day)	Difference with respect to the base case (day)	
	FNR for accumulation days (%)*			
	FNR for melt days (%)**			

The day of peak SWE is compared to its observed value by a difference in days.

Two additional performance metrics measure how accurately the model predicts the timing of melt and accumulation events. The annual false-negative rate (FNR) for accumulation is the percent of days when an observed accumulation event occurs but a modeled one does not, such that

$$\mathrm{FNR}_{\mathrm{acc}} = 100\% \times \mathrm{FN}/P, \tag{3}$$

where FN is the number of days over the water year that shows observed accumulation without modeled accumulation and P is the number of days with observed accumulation. A higher $\mathrm{FNR}_{\mathrm{acc}}$ means that there are more days in which the model fails to simulate an observed accumulation event. The $\mathrm{FNR}_{\mathrm{melt}}$ similarly measures how often the model fails to predict an observed daily melt event. For both FNRs, interannual values are averaged into a single metric for each station. Note that, as opposed to the daily melt and storm rate metrics, these FNRs are not constrained by minimum melt or accumulation thresholds.

c. *Evaluation of model sensitivity*

We assessed the sensitivity of the modeled snow metrics for each model experiment relative to the base case in two ways (Table 2). First, we identified where significant sensitivities exist by applying the Kolmogorov–Smirnov (KS) test (Text S6) to each snow metric at all stations in each experiment. This test has been utilized to assess model sensitivity in numerous hydrologic modeling studies (e.g., He et al. 2011; Sun et al. 2019). A two-sample KS test was performed at each station to test whether the snow metrics from the base case and the experiment come from the same distribution. KS values range from 0 to 1, with higher values indicating greater sensitivities (changes). We used a minimum KS threshold value of 0.5 to identify sensitivity because it yields statistically significant results at p value < 0.1. Stations with a KS statistic equal to or greater than 0.5 were considered sensitive to that experiment for that snow metric.

Second, we quantified the magnitude (and direction) of changes in snow predictions by comparing each model experiment's simulated snow metrics to those from the base case by applying Eq. (1), but with the base case as the reference. Changes in snow metrics for each experiment are thus expressed as the percent change in metric value relative to the base case.

5. Results

a. *Model performance relative to SNOTEL observations*

1) Base case performance across metrics

Overall, the Noah-MP base case overestimates accumulated SWE observations at 90% of sites, with an overall median overestimate of 9.6% (Fig. 3a). This overestimation most likely results from excessive snowfall input due to inaccurate precipitation partitioning. This overestimation is consistent with findings by Letcher et al. (2022) over New York State; they also identify the precipitation partitioning scheme as a source of significant sensitivity in Noah-MP. Stations with underestimated accumulated SWE have a median bias of only −1.7%. The observed storm rate is underestimated by the model in 53% of stations, which have a median bias of −3.2% (relative to the overall median of −0.4%; Fig. 3b). The median FNR is 8% for accumulation days, meaning that the model correctly simulates accumulation on 92% of observed days (Fig. 3c). Peak SWE occurs later in the base case than observations at 74% of stations, for which the median bias is 9.5 days (relative to an overall median of 5.4 days; Fig. 3d). Stations where the model predicts earlier peak SWE have a median bias of only −4 days. In general, model performance during the melt season is worse than in the accumulation season. The daily melt rate is underestimated by the model at 79% of stations, by an overall median bias of −15% (Fig. 3e). The median FNR for melt is 38%, which is over 4 times greater than that for accumulation (Fig. 3f). This means that the model agreement with the timing of observed melt events is much worse than that for accumulation.

Model performance varies across geographic regions. For example, stations in the Cascades region exhibit a statistically different model performance ($p < 0.05$) than other regions across all snow metrics. No stations in the Cascades nor the Sierra Nevada underestimate accumulated SWE. The Cascades has the highest median bias in accumulated SWE (19.5%, 2.5 times greater than the median of all other stations) (Fig. 3a), indicating lower model accuracy during the accumulation season. Of note, the Sierra Nevada, Cascades, and Arizona/New Mexico mountains are the only regions where the median bias in storm rate is positive, suggesting that the model tends to overestimate accumulation on large storm days (Fig. 3b). Inaccurate precipitation partitioning could cause the model to estimate more SWE accumulation than observed on large storm days. This is notable because large storms like those produced by atmospheric rivers dominate snowpack accumulation in the Pacific mountainous regions, particularly in the Sierra Nevada (Serreze et al. 2001). Generally, stations in the interior ranges show a better performance in the accumulation season than those in the coastal ranges: four out of the five interior regions (Blue Mountains, Basin and Range, northern Rockies, Wasatch and Uinta Mountains, and southern Rockies) have a median bias below the overall median in both accumulated SWE and FNR for accumulation (Figs. 3a,c). For the melt season, the Cascades region has the lowest median bias in timing of peak SWE (1.3 days), the smallest median bias in daily melt (−2.4%), and the lowest FNR for melt days (30%) (Figs. 3d–f). This means that the model performs significantly more accurately over the melt season in the Cascades than in other regions. The southern Rockies also have a statistically different melt season performance, with the largest bias in timing of peak SWE (median delay of 21 days) (Fig. 3d), the most negative median bias in simulated daily melt rate (−27%) (Fig. 3e), and the second-highest median FNR for melt days (51%) (Fig. 3f). This suggests that melt season model performance is the poorest in the southern Rockies. It is worth noting that the Cascades stations are the snowiest and second warmest (average observed accumulated SWE of 796 mm and

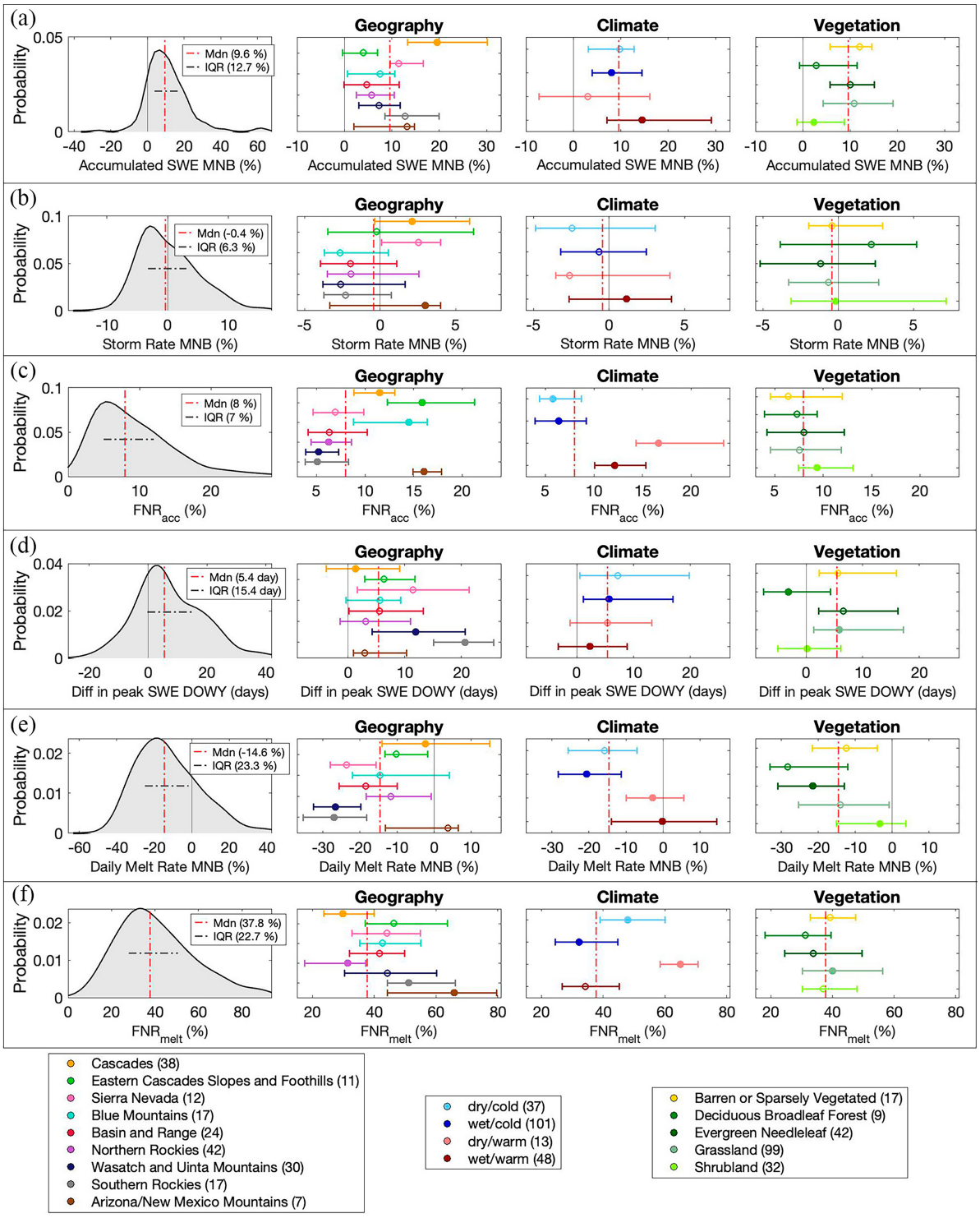

FIG. 3. Model performance across all stations in the base case, as compared to SNOTEL SWE observations. Six performance metrics were used: (a) MNB in accumulated SWE; (b) MNB in storm rate; (c) FNR for accumulation days; (d) difference in timing of peak SWE; (e) MNB in daily melt rate; and (f) FNR for melt days. (first column) A smoothed histogram of the performance metrics across stations in the base case. A vertical dashed red line indicates the median metric value, and a horizontal dashed gray line indicates the interquartile range (IQR). The performance metrics are separated (second column) by geographic region, (third column) by climate subgroup, and (fourth column) by vegetation type. Circles mark the median of the subgroup, and the width of the line marks the IQR. If the subgroup has a filled-in circle, it is considered significantly different (p value < 0.05) from the other subgroups. The number of stations in each subgroup is noted in the legend entries.

winter temperature of −0.12°C; Table S2), and the southern Rockies stations are the coldest (average observed winter temperature of −5.3°C; Table S2). Also relevant here is the role of light-absorbing particles (i.e., dust and soot) on snow: Albedo decay and shortened snow cover duration caused by dust on snow have been widely observed in the Rocky Mountains and Wasatch and Uinta Range (e.g., Painter et al. 2012, 2007). Noah-MP does not explicitly capture radiative forcing by dust on snow and so can be expected to underestimate the observed melt rates in regions impacted by dust deposition. This is consistent with the result that melt rate bias is most negative in the Wasatch and Uinta Mountains and southern Rockies (Fig. 3e). Note that the alternative BATS experiments use a parameter value optimized for sites with considerable dust on snow effects; median melt rate bias is still negative and largest in the Wasatch and Uinta Mountains and southern Rockies in these experiments, but the error magnitude in $BATS_{tau\_NIR}$ ($BATS_{tau\_VIS}$) decreases (increases) at these sites by a median of 14% (16%) relative to the base case (Fig. S7).

Variations in model performance can also be explained by differences in climate. Sites classified as cold (dry/cold or wet/cold) demonstrate similar bias levels in accumulated SWE regardless of precipitation amount, whereas warm sites (dry/warm or wet/warm) show more bias (on median 4.7 times higher) when they are also wet (Fig. 3a). This suggests that precipitation at warm sites is a more distinguishing factor for accumulated SWE bias than at cold sites. Dry sites (dry/cold or dry/warm) have a more negative bias (−2.5%) in storm rate than wet sites (−0.2%) (Fig. 3b), indicating that the model better predicts the storm rate when precipitation is higher (and storms are typically larger). Temperature instead is the distinguishing factor for FNR: On median, warm sites have an accumulation FNR twice as high as that for cold sites (Fig. 3c). So, the model predicts accumulation events better at colder stations, where precipitation partitioning thresholds are less influential in determining snowfall. In the melt season, dry/cold stations show the highest bias in timing of peak SWE and wet/warm stations show the least (Fig. 3d). Temperature distinguishes model performance in daily melt rate: Cold sites have a median bias 50 times greater (−20%) than warm sites (−0.4%) (Fig. 3e). FNR for melt days is instead more distinguishable by precipitation: Dry stations have a median bias 1.6 times higher than wet stations (Fig. 3f). Both wet/warm sites and wet/cold sites perform statistically differently than other climate subgroups in five of the six metrics (Fig. 3), suggesting that precipitation amount strongly affects model performance.

Vegetation type also plays a role in model performance, as the amount and type of vegetation can impact both energy and mass balance of snow. Note that vegetation type is manually set in the model to match the SNOTEL station; 50% of stations are located in clearings and labeled as "grassland." It is worth mentioning that at more vegetated sites, snow simulation becomes more complex. Although the Noah-MP canopy model correctly accounts for shadowing effects on solar radiation, it could be introducing an inconsistency between modeled SWE and observed SWE because it models canopy snow interception, whereas the SNOTEL snow pillows are likely less affected by canopy interception because they are not typically placed directly under trees even when the site is forested. This could lead to an underestimation of observed SWE at forested sites in this study by Noah-MP.

Shrubland sites show statistically different model performance at most metrics (Figs. 3a–e), signifying a unique model performance. These sites exhibit the lowest median bias magnitude in accumulated SWE (2.4%) and storm rate (−0.2%), but the highest bias in FNR for accumulation days (9%) (Figs. 3a–c). This suggests that the model simulates snow volume relatively well at shrubland sites, but does so more poorly for the timing of accumulation events, perhaps because of vegetation interception. Deciduous broadleaf sites have a bias in accumulated SWE less than the overall median and are the only ones to have a negative bias in timing of peak SWE (−3 days) indicating earlier simulated snowmelt onset (Fig. 3d). In the melt season, the model simulates median melt rate 1.7 times more accurately at nonforested sites than forested sites (Fig. 3e), but FNR 1.8 times more accurately at forested sites (Fig. 3f). This means that at forested sites, the model simulates the daily melt rate less accurately, but those melt events are more likely to occur on the correct (observed) days. It should be noted that patterns in model performance across vegetation types may be cross correlated with climate, geography, and other factors.

2) Relationship between base case performance and climate variables

Studying the correlations between model error and climate-related variables underscores climate as a crucial factor affecting model performance. Figure 4 presents correlations R between the observed winter temperature, winter precipitation, snow/precipitation ratio (peak SWE/annual precipitation), and accumulated SWE, and base case model error across all sites. Model error is defined here as mean absolute error, absolute difference, or FNR (Table 2). By using the absolute error rather than bias, these correlations reveal the relationship between climate variables and error magnitude irrespective of error direction.

Over the accumulation season, warmer winter temperatures correspond to higher error in accumulated SWE ($R = 0.39$), storm rate ($R = 0.31$), and FNR ($R = 0.55$). This supports the finding from Fig. 3 that warm stations perform more poorly during the accumulation season than cold stations, likely because these sites are more sensitive to inaccurate precipitation partitioning. Winter precipitation also positively correlates with model error in accumulated SWE ($R = 0.28$) (Fig. 4). This reaffirms the finding in Fig. 3 that wet/warm sites have the poorest model performance in accumulated SWE. Snow/precipitation ratio strongly negatively correlates with model error in accumulated SWE ($R = -0.55$), storm rate ($R = -0.38$), and FNR ($R = -0.59$) (Fig. 4). This suggests that sites with a lower snow/precipitation ratio (which tend to have warmer temperatures) have higher accumulation errors. This is consistent with the finding that observed accumulated SWE negatively correlates with error in storm rate ($R = -0.3$) and FNR

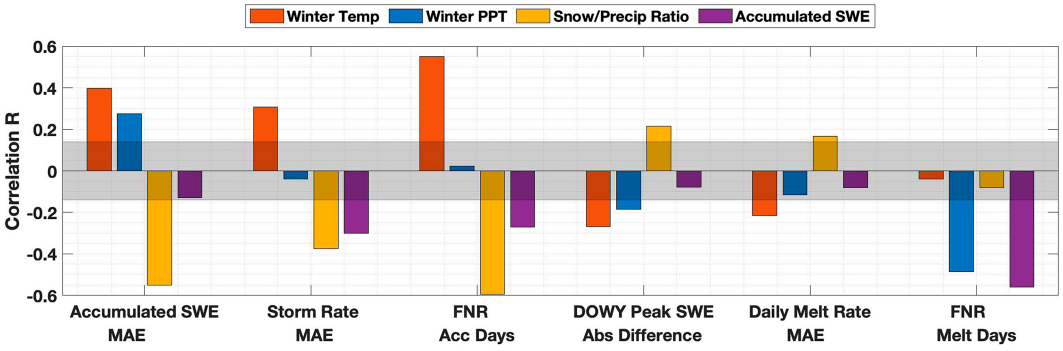

FIG. 4. Correlations R between seasonal climate variables and model error across stations in the base case as represented by six performance metrics: MAE in accumulated SWE, MAE in storm rate, FNR for accumulation days, absolute difference in timing of peak SWE, MAE in daily melt rate, and FNR for melt days (Table 2). Climate-related variables are computed for each station year from observed records over winters (November–March) of WYs 2007–19. The gray area marks R values that are not statistically significant (p value > 0.05). Positive (negative) bars indicate that a higher value for the climate variable correlates with higher (lower) model error.

($R = -0.27$), indicating that the model is worse at predicting accumulation events at less snowy sites.

Generally, the trends between climate-related variables and model error are opposite and weaker for melt season metrics. Winter temperature negatively correlates with error in timing of peak SWE ($R = -0.26$) and daily melt rate ($R = -0.21$) (Fig. 4), indicating that colder stations perform worse than warm stations in the melt season. The strong negative correlation between winter precipitation and FNR for melt days ($R = -0.49$) (Fig. 4) reinforces the finding from Fig. 3 that wet sites perform distinctively worse than dry sites for this metric. This is also true, but less significant, for the relationship between winter precipitation and error in timing of peak SWE ($R = -0.19$) (Fig. 4). Observed accumulated SWE has a strong negative correlation with FNR for melt days ($R = -0.56$) (Fig. 4), indicating that the model more often fails to predict observed melt events at sites with less snow. Note that observed accumulated SWE is the only variable which relates negatively to model error across both accumulation and melt seasons; so, less snowy sites consistently perform worse than snowier sites. The positive correlations between snow/precipitation ratio and error in timing of peak SWE ($R = 0.22$) and daily melt rate ($R = 0.17$) are less significant but demonstrate poorer model performance in the melt season at stations with a high snow/precipitation ratio (which also tend to be colder).

b. *Sensitivity of snow simulations to model configurations*

1) IDENTIFYING SENSITIVITY TO CHANGES IN MODEL CONFIGURATION

To assess whether changes in snow metrics caused by the model experiments amount to a significant model sensitivity, we applied the KS test to four snow metrics (Table 2). Snow metrics at a station are considered significantly sensitive to a particular experiment if the KS statistic is equal to or greater than 0.5 (p value < 0.1). Figure 5 summarizes how many stations are considered sensitive to each experiment and snow metric.

Overall, only two experiments show significant sensitivity in accumulation metrics at over 5% of stations: AORC and Precip0 (Fig. 5). Precip0 lowers the temperature threshold for snowfall to 0°C, leading to a sensitivity in accumulated SWE and storm rate at 22% and 4% of stations, respectively (Fig. 5). In contrast, the other precipitation partitioning experiment Precip2.2, which raises the partitioning threshold to 2.2°C, has at most 1% of stations with accumulation metric

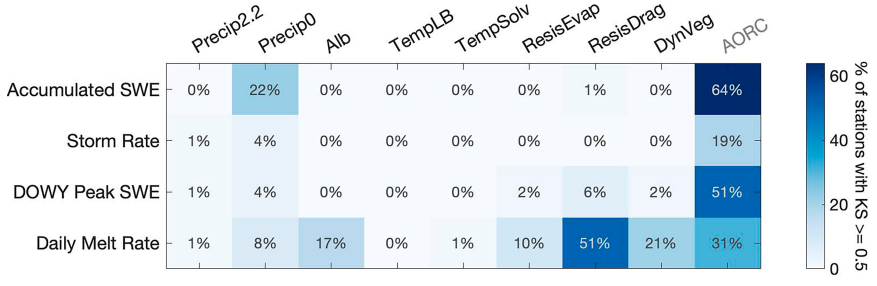

FIG. 5. The percent of sensitive stations (KS statistic ≥ 0.5, p value < 0.1) in each experiment and across four snow metrics. Note that AORC is an experiment that tests forcing input rather than model physics.

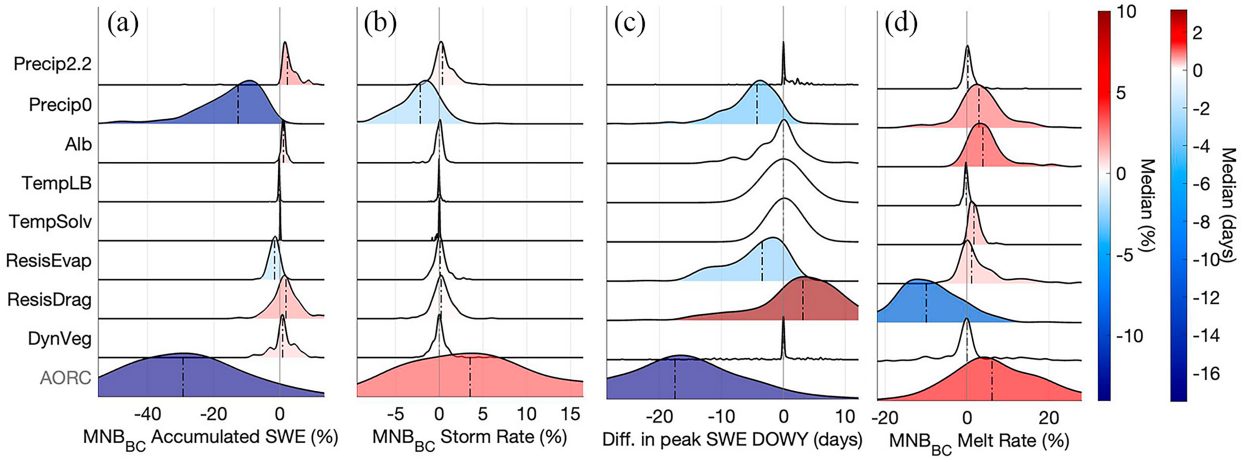

FIG. 6. Distributions of changes in snow metrics relative to the base case across SNOTEL stations and nine experiments for four snow metrics. (a) Accumulated SWE, (b) storm rate, (c) timing of peak SWE, and (d) daily melt rate. Bias metrics (MNB_{BC} or difference; Table 2) are computed for each station year with reference to the base case and then averaged for each station over the time period. The distribution color and the dashed horizontal line correspond to the median bias value for each experiment. A red (blue) color indicates the model configuration produces a higher (lower) median value than the base case.

sensitivity (Fig. 5), indicating that the 0°C threshold yields significantly more deviation in results from the base case (which uses a linear increase in partitioning with temperature per Jordan 1991) than 2.2°C. The AORC experiment, which changes the input precipitation and air temperature, demonstrates the highest levels of sensitivity, with 64%, 19%, and 51% of stations sensitive to accumulated SWE, storm rate, and timing of peak SWE, respectively (Fig. 5).

Six experiments exhibit significant sensitivity in melt metrics at over 5% of stations: Alb, Precip0, ResisEvap, ResisDrag, DynVeg, and AORC, indicating that melt metrics are sensitive to more changes in model configuration than accumulation metrics. Alb and DynVeg both impact surface albedo, resulting in 17% and 21% of stations showing melt rate sensitivity, respectively. ResisDrag, which impacts the computation of surface heat fluxes, is the physics-related experiment with the most sensitivity in the melt season: 51% of stations have melt rate sensitivity and 6% have sensitivity to timing of peak SWE. ResisEvap also affects surface heat fluxes and demonstrates melt rate sensitivity at 10% of stations. The Precip0 experiment exhibits sensitivity to daily melt rate in 8% of stations; this is an indirect effect of changes to accumulation patterns. At most 1% of stations show sensitivity to any metric in TempLB and TempSolv, meaning that these model configurations do not deviate greatly from the base case. AORC shows 31% of stations with sensitivity in melt rate (Fig. 5).

2) CHANGES IN SNOW METRICS RELATIVE TO BASE CASE ACROSS MODEL CONFIGURATIONS

The magnitude and direction of changes to snow predictions caused by alterations to model configuration or forcing input are evaluated by comparing four snow metrics in each experiment to base case results (Table 2). Figure 6 illustrates the distributions of these bias metrics for each experiment.

The precipitation partitioning experiments (Precip2.2 and Precip0) are expected to primarily impact snowfall and snow accumulation. Precip2.2 increases the temperature threshold for precipitation partitioning, resulting in a slightly higher accumulated SWE (median of 2.3%) (Fig. 6a), but less notable impacts to storm rate (median of 0.2%) and timing of peak SWE (median of 0.2 days) (Figs. 6b,c). Precip0 instead lowers the temperature threshold, leading to less snowfall than the base case. The storm rate and total accumulated SWE notably reduce as a result, by a median of 4.3% and 13%, respectively (Figs. 6a,b). The day of peak SWE is correspondingly advanced by a median of 4.3 days (Fig. 6c). The range of bias for accumulated SWE in Precip0 is the widest for accumulation metrics [interquartile range (IQR) of 12.3%], indicating that the effect of this experiment varies greatly across stations [see section 5b(3)] (Fig. 6a).

The subsequent five experiments relate primarily to melt processes. Alb generates a higher daily melt rate (median of 4%) than the base case (Fig. 6d). This is consistent with the observation that the alternative albedo option generates lower snow albedo (Niu et al. 2011; Abolafia-Rosenzweig et al. 2022). ResisEvap, which alters surface resistance to turbulent fluxes, generates a slightly higher melt rate (median of 1.1%) (Fig. 6d). This implies that the change in surface resistance was enough to reduce latent heat flux so that the additional energy at the surface initiated higher snowmelt rates. ResisDrag, which alters how the surface drag coefficient is estimated, impacts melt metrics more significantly. The alternative option (Chen et al. 1997) has been noted to produce a lower value for the surface drag coefficient (e.g., Zhang et al. 2014), resulting in higher heat flux values. This is consistent with the experiment reducing daily melt rate, by a median of 9.9% (Fig. 6d). This experiment has a wide range of bias values for daily melt rate (IQR of 10%), suggesting that the impact of ResisDrag varies notably across sites [see section 5b(3)]. TempSolv, which

changes the solver approach for the soil temperature equation, increases the daily melt rate by a median of 1.7% (Fig. 6d). TempLB, which changes the temperature lower boundary condition of the soil column, generates insignificant to no change (less than 0.2%) to all metrics (Figs. 6a–d).

DynVeg and AORC are expected to impact both accumulation and melt processes. The AORC experiment uses the same physics configuration as the base case, but with different precipitation and temperature forcings. AORC total winter precipitation differs from SNOTEL observations heterogeneously across the WUS (Fig. S1a) but averages out to 10.6% less. AORC winter temperature underestimates observations by an average of 0.2°C (Fig. S1b). As a result of this and/or differences in individual storm events, the storm rate increases in the AORC experiment (median of 3.5%) (Fig. 6b). Peak SWE occurs at a median of 17.5 days earlier, which, combined with a higher melt rate (median of 6.1%; Fig. 6d) and lower input precipitation (Fig. S1a), is consistent with total accumulated SWE decreasing (median of −29.3%) (Fig. 6a). The range of bias values in AORC is wider than in any other experiment (IQRs ranging 11%–25% across metrics), suggesting that snow simulation varies greatly across sites in this experiment and that the experiment introduces more uncertainty (Fig. 6). DynVeg turns on the dynamic vegetation module. As a result, accumulated SWE changes slightly (median of 1%). There is insignificant to no change in median values (less than 0.2%) for peak SWE timing, storm rate, and melt rates (Figs. 6a,b,d). Of note, the IQR for the melt rate biases is 5% (Fig. 6d), indicating that DynVeg impacts melt processes at some stations more significantly than the median suggests.

Figure 6 also illustrates changes in snow simulation due to indirect effects by the model experiments. For example, Precip0, which directly reduces snowfall relative to the base case, also exhibits a slight increase in daily melt rate (median of 3%) (Fig. 6d). This is because a shallower snowpack presumably reaches the melt phase earlier. Although ResisDrag primarily impacts melt rate, it also demonstrates a slight increase in accumulated SWE (median of 1.9%) and a corresponding delay in timing of peak SWE (median of 3.2 days) (Figs. 6a,c). This is consistent with less winter melt occurring due to lower melt rates. The IQR for biases in timing of peak SWE in this experiment is the second widest (7.5%), indicating a varied impact that can result in either positive or negative biases at individual sites. ResisEvap shows effects similar in magnitude but opposite in direction: a slight decrease in accumulated SWE (median of −1.6%) and advance in timing of peak SWE (median of −3.4 days) (Figs. 6a,c). This is consistent with more winter melt occurring due to a higher melt rate.

3) Changes in snow metrics relative to base case across site conditions

Changes in snow metrics caused by different model configurations relative to the base case can also be evaluated across site conditions in order to understand how and why snow simulation responds differently to alterations in physics processes or forcing input. Only the metrics with the highest percent of stations showing a significant KS statistic (Fig. 5) are evaluated in this section.

Although the AORC experiment shows high levels of sensitivity (Fig. 5), the changes to snow metrics caused by this experiment relate to the localized differences between SNOTEL and AORC temperature/precipitation rather than model physics and so are more heterogeneous across geography, climate, and vegetation than in other experiments.

Precip0 primarily impacts accumulated SWE as it changes the amount of snowfall input (Fig. 5). Figure 7a illustrates how stations closer to the Pacific coast—in the Cascades, Eastern Cascades Slopes and Foothills, and the Sierra Nevada—show a more negative and larger change in accumulated SWE than stations located in the interior ranges—northern Rockies, Blue Mountains, Basin and Range, and Wasatch and Uinta Mountains. This is due to more temperate climate in the more coastal areas, with warmer temperatures that hover more closely to the precipitation partitioning threshold. In fact, stations classified as warm (either dry/warm or wet/warm) demonstrate a higher and more negative median bias in snow accumulation than stations classified as cold (Fig. 7a). Precipitation amount also matters: Within each temperature grouping (warm or cold), stations classified as wet demonstrate a higher bias than their dry counterparts (Fig. 7a). So, Precip0 causes more change in accumulated SWE in wetter and warmer stations because changing the precipitation partitioning threshold has a bigger impact where winter temperature is closer to that threshold.

Alb primarily impacts the daily melt rate (Fig. 5), and that impact varies by geography and climate (Fig. 7b). The only region to show a negative change in daily melt rate in this experiment is the Arizona/New Mexico mountains (Fig. 7b), which is the least snowy region in the study (Table S2). The region with the highest positive change in daily melt rate (and widest IQR) is the southern Rockies (Fig. 7b). Both of these regions are statistically different than the others. Precipitation amount rather than temperature is a distinguishing factor for this experiment: Stations classified as wet (either wet/cold or wet/warm) have a higher median bias in melt rate than stations classified as dry (Fig. 7b). Dry/warm stations have the lowest median bias, whereas wet/cold stations have the highest (Fig. 7b).

ResisEvap and ResisDrag also primarily impact daily melt rate (Fig. 5). For ResisEvap, stations in the eastern Cascades and Foothills demonstrate the most negative and highest absolute median bias of all regions (Fig. 7c). The only other region with an absolute median bias higher than 2% is the southern Rockies, which has the highest positive bias (Fig. 7c). In this experiment, the IQRs of the bias metrics vary notably across regions—for example, the IQR for the dry Arizona/New Mexico region is much wider than the wet Cascades region (Fig. 7c). While the median biases for each climate subgroup do not differ significantly, the IQRs do. Stations classified as dry (either dry/cold or dry/warm) have a wider IQR than stations classified as wet (Fig. 7c). So, the changes in daily melt rate caused by ResisEvap range wider across drier stations.

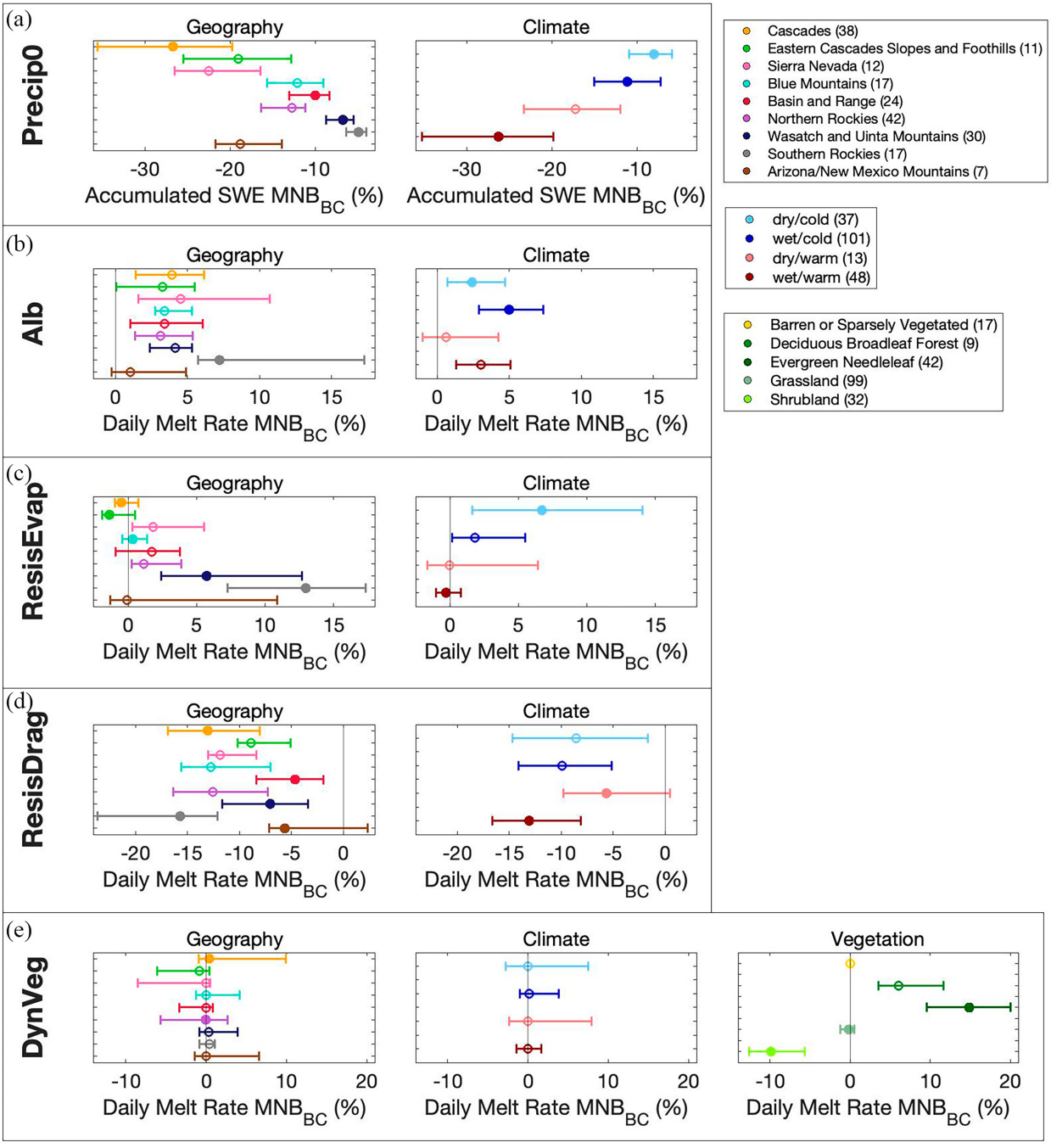

FIG. 7. Distribution of changes from the base case in key snow metrics across site conditions in (a) Precip0, (b) Alb, (c) ResisEvap, (d) ResisDrag, and (e) DynVeg experiments. The bias metric MNB relative to base case for accumulated SWE is shown for Precip0 and for daily melt rate for the other experiments. The first column of subpanels presents results across different geographic regions of the WUS, and the second column separates results by climate subgroups. Results are separated by vegetation type for the DynVeg experiment. Circles mark the median of subgroup, and the width of the line marks the IQR. If a subgroup has a filled-in circle, it is considered significantly different (p value < 0.05) from the other subgroups. The number of stations in each subgroup is noted in the legend entries.

Precipitation amount is also a distinguishing factor for changes to the daily melt rate in ResisDrag. Wet stations have a more negative bias than dry stations, with wet/warm stations demonstrating the highest absolute bias (Fig. 7d). Geographically, the Cascades demonstrate the most negative and highest absolute median bias, whereas the Arizona/New Mexico mountain region shows the lowest median bias but with the widest IQR (Fig. 7d).

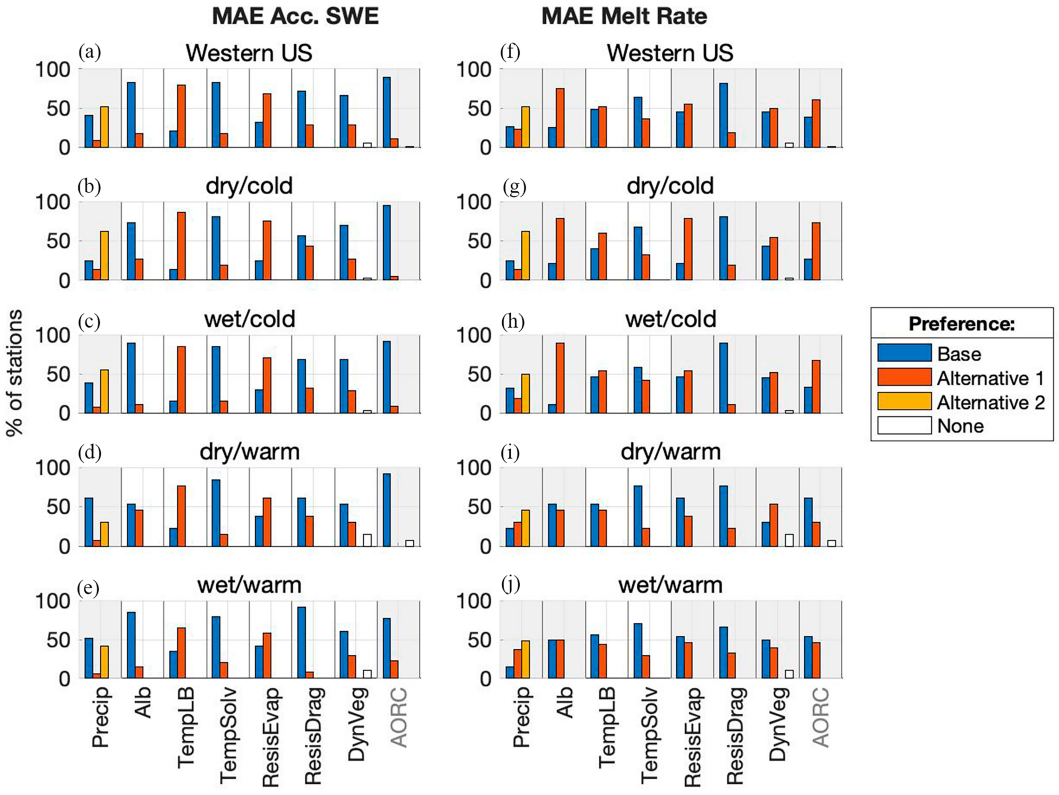

FIG. 8. Fraction of stations showing better model performance, as represented by (a)–(e) the lowest MAE in accumulated SWE and (f)–(j) MAE in daily melt rate, for each option under seven physics processes and one pair of forcings input, for all stations, and for subsets covering four climate categories. Note that for the precipitation partitioning process, Precip2.2 is "alternative 1" and Precip0 is "alternative 2." There are 37 stations in the dry/cold subgroup; 13 in dry/warm; 101 in wet/cold; and 48 in wet/warm. If a station shows no difference in model performance between the different options, it is added to the white bar. The best-performing option has the highest bar. The gray background highlights the most relevant options for that snow metric by indicating which experiment causes sensitivity relative to the base case for that snow metric at over 5% of stations (as shown in Fig. 5).

For DynVeg, more stations show sensitivity to daily melt rate than any other snow metric (Fig. 5). Although the median biases across geographic regions and climate subgroups are similar and close to 0, there are notable differences in the IQRs (Fig. 7e). For example, stations in the Cascades region have the widest IQR, whereas stations in the southern Rockies have the narrowest IQR (Fig. 7e). Across climate subgroups, the IQR is wider in dry stations (either dry/cold or dry/warm) than wet stations (Fig. 7e). These differences suggest that station groups with wider IQRs have more uncertainty in how DynVeg affects snow simulation. Changes to melt rate caused by DynVeg are notably discernable by vegetation type. There is no change from the base case daily melt rate at barren or sparsely vegetated stations, suggesting that the dynamic vegetation module (see more in Text S4) does not have any impact on snow simulation at these sites. On the other hand, there are negative change for shrubland stations and positive change for forested stations (either deciduous broadleaf or evergreen needleleaf) (Fig. 7e). Grassland stations show a median bias close to 0, with individual stations showing slight changes (Fig. 7e).

c. *Best-performing physics options*

The variations in model performance and sensitivity across sites speak to inadequacies in physical process representation that manifest differently under different site conditions. This implies that customizing model configurations to the unique characteristics of the model domain, particularly climate type which we have found has strong patterns with model behavior, could yield more accurate and useful results. Figure 8 illustrates which option—either the base case or alternative—performs best relative to observations, as measured by the MAE in accumulated SWE and daily melt rate (Table 2).

Overall, the Precip0 alternative for precipitation partitioning yields the best accumulation season model performance at most stations across the WUS (Fig. 8a). This option reduces snowfall amount relative to the base case and partly compensates for the positive bias in accumulated SWE in the base case (Fig. 3a). At most warm stations, which are more sensitive to changes in temperature thresholds for precipitation partitioning (Fig. 7a), the base option instead more accurately predicts accumulated SWE.

As expected, using observation-based forcings rather than gridded AORC forcings leads to more accurate accumulated

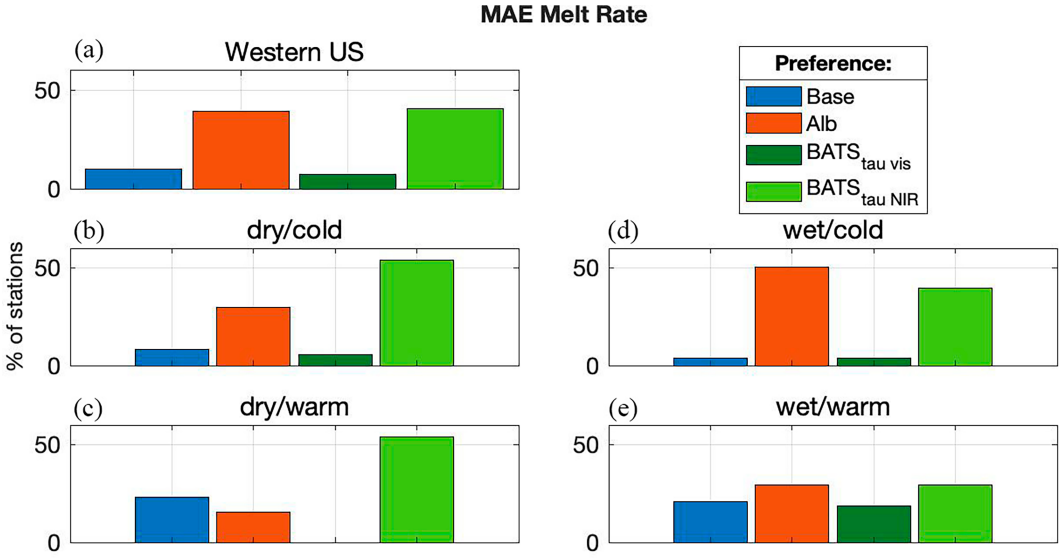

FIG. 9. Fraction of stations showing better model performance, as represented by the lowest MAE in daily melt rate, for four snow albedo experiments: BATS with default parameters (base case), CLASS (Alb), BATS with tuned parameter ($BATS_{tau\_vis}$), and BATS with tuned parameter ($BATS_{tau\_NIR}$) (Table 1), (a) for all stations and for subsets covering four climate categories: (b) dry/cold, (c) dry/warm, (d) wet/cold, and (e) wet/warm. The best-performing option has the highest bar.

SWE at most stations (Figs. 8a–d). At those few stations where AORC instead performs better, the base case model configuration can be considered erroneous because the AORC forcings and validation dataset (observed SWE) are inconsistent, and thus, the AORC experiment is not expected to yield superior model performance than the experiment which uses observation-based forcings.

At predicting the observed melt rate, the alternative for Alb vastly outperforms the base option at most stations across the WUS and especially at cold stations (Figs. 8e–g). This is consistent with the finding that observed melt rate is generally underestimated in the base case (Fig. 3e), and the alternative snow albedo option yields lower albedo values and thus a higher melt rate (Fig. 6d). At warm stations, where melt rates are typically higher and the change relative to the base case is less (Fig. 7b), the base option instead outperforms the alternative at most stations. Note that in this case, the base option utilizes all default parameters for the BATS albedo scheme. Figure 8 illustrates how tuning one parameter in this scheme to a locally optimized value (Abolafia-Rosenzweig et al. 2022) affects the results.

Similarly, the alternative for ResisEvap, which tends to increase the melt rate relative to the base case (Fig. 6d), outperforms the base option at most stations across the WUS (and in particular cold stations). Conversely, the base option outperforms the alternative for ResisDrag, which reduces the melt rate relative to the base case at shrubland and grassland sites (Fig. 7e), at most stations across all climate groups except for wet/warm (Figs. 8f–j). Note that the impact of the DynVeg alternative relative to the base case is more correlated with vegetation type than climate subgroup (Fig. 7e). The second alternative for precipitation partitioning (Precip0), which indirectly impacts melt rate by reducing accumulated snowfall, outperforms the base option across all climate groups (Figs. 8f–j). Also, using gridded forcings (AORC) yields more accurate melt rates than using observation-based forcings at most stations except for those in warm climates (Figs. 8f–j), suggesting that the base case configuration is not optimized for melt season SWE predictions. Notably, for most stations where AORC outperforms the base option, the alternative snow albedo model also outperforms the base option (Figs. 8f–h)—this suggests that a key source of error in the base case configuration is the use of an inadequate snow albedo model (in the base case, BATS with default parameter values) especially at cold stations.

Note also that the alternative option sometimes outperforms the base case even among the physics processes that do not yield significant changes relative to the base case. This occurs in the accumulation season at most stations for TempLB and ResisEvap and in the melt season at most cold stations for TempLB.

The BATS snow albedo model, used in the base case and the default within Noah-MP, has 12 tunable parameters, some of which have been found to significantly affect the performance of the model (e.g., Abolafia-Rosenzweig et al. 2022). Figure 9 illustrates how tuning one of those parameters τ_0 affects the model's ability to reproduce observed melt rate. Overall and for dry stations, the $BATS_{tau\_NIR}$ experiment outperforms the other snow albedo configurations, yielding the lowest melt rate error in 41% of all stations (Figs. 9a–c). This demonstrates that tuning even one parameter of the BATS model, such as to locally optimized values, can improve the model performance such that it exceeds that of an alternative snow albedo scheme. Specifically, decreasing the τ_0 parameter

value (as in $BATS_{tau\_NIR}$) improves the ability of the BATS model to reproduce observed melt rate at most stations. See Fig. S6 for distributions of melt rate bias in $BATS_{tau\_NIR}$ and $BATS_{tau\_vis}$. Conversely, Alb (which uses the CLASS albedo scheme) outperforms the BATS experiments at most wet stations (Figs. 9d,e). Note that the adjusted values for the τ_0 parameter were optimized to local observations in the southern Rockies (Abolafia-Rosenzweig et al. 2022). The results summarized in Fig. 9 suggest that these locally optimized values can effectively improve model performance at other sites with similar climates (i.e., dry/cold; Fig. 9b), but less effectively at dissimilar sites (i.e., wet/warm; Fig. 9e).

6. Conclusions

This study examines snow water equivalent (SWE) simulation by Noah-MP across 199 sites in the western United States. The base case and eight model configuration experiments test physics options related to precipitation partitioning (Precip2.2 and Precip0), snow albedo (Alb), lower soil temperature boundary condition (TempLB), snow/soil temperature time scheme (TempSolv), surface resistance to evaporation/sublimation (ResisEvap), surface layer drag coefficient (ResisDrag), and dynamic vegetation (DynVeg). These experiments are forced by in situ meteorology in order to rigorously identify model deficiencies. The ninth experiment (AORC) tests an alternative source of forcings to provide insight into forcing errors and relative model uncertainties.

With respect to how well Noah-MP can reproduce SWE at sites across the WUS, we find that the base case, which matches the National Water Model (NWM) configuration and uses observed temperature and precipitation forcings, overestimates observed accumulated SWE at 90% of stations by a median of 9.6%. Inaccurate precipitation partitioning by the model could explain these errors. At most sites, it also predicts later peak SWE timing (5.4 days) and underestimates daily melt rate (median of −14.6%). The use of globally applied default parameters in the base case albedo model (BATS) could explain this underestimation. The model more successfully predicts the timing of observed accumulation events than observed melt events.

With respect to the sensitivity of the model to alternative configurations, we find that Precip0 and AORC demonstrate significant sensitivity in the accumulation season, while Alb, Precip0, ResisEvap, ResisDrag, DynVeg, and AORC exhibit significant melt season sensitivity. Precip2.2, TempSolv, and TempLB show little to no significant sensitivity. Of the model configurations that test physics processes, the greatest change to accumulation season predictions occurs in Precip0 (median of −13%). On average, Precip0 reduces accumulated SWE, storm rate, and timing of peak SWE, while Precip2.2 slightly increases accumulated SWE. With regard to melt season predictions, the greatest change occurs in ResisDrag (median of −10%). ResisDrag on average reduces melt rate, increases accumulated SWE, and delays peak SWE timing, while ResisEvap does the opposite. DynVeg demonstrates a wide range of changes to melt rate, which are discernable by vegetation type but average out to a near-zero effect overall. TempLB and TempSolv have minimal effects. AORC causes substantial changes in snow performance metrics that often exceed those caused by alterations in model physics in both magnitude and range (medians ranging from −29% to 6%). This suggests that previous assessments of model performance in Noah-MP or other land surface models (LSMs) could be masked by forcing errors or inconsistencies with the validation dataset and emphasizes that generating accurate meteorological forcings should be prioritized over adjusting model physics representation for promoting more accurate SWE simulation in LSMs.

We also find that the model's performance relative to observations and sensitivity relative to the base case differs across regions, climates, and vegetation types, with the strongest trends related to climate. The model performs better relative to observations in the accumulation season at colder sites with a higher snow/precipitation ratio (such as in the southern Rockies) and in the melt season at warmer stations with a lower snow/precipitation ratio (such as in the Cascades). The model performs most uniquely in the Cascades region—perhaps because these sites are warmer despite their deep snowpack. We also find that not all stations share the same best-performing model configuration, highlighting inconsistencies in how the model simulates SWE and suggesting the need for models customized to site conditions. Notably, at most stations and especially those classified as cold, the Precip0 alternative outperforms the base case for predicting accumulated SWE; the Alb, ResisEvap, and DynVeg alternatives outperform the base case for predicting melt rate. We find that tuning just one parameter of the BATS snow albedo model yields lower melt rate errors at most stations. These findings indicate that the current default NWM configuration is not optimized for all sites of the western United States.

Further research should explore how the snow model responds to superposing changes in model configurations and consider the implications of using varying configurations for different site conditions such as climate when running large-scale models like the National Water Model.

Acknowledgments. The author would like to thank Dr. Dennis Lettenmeier, who provided the initial ideation and funding for this paper under NOAA OAR/OWAQ Award NA18OAR4590396, along with project co-I's Dr. Kostas Andreadis and Dr. Steve Margulis; and Dr. Lu Su, who provided key guidance to set up and run the model. We would also like to thank the three reviewers and editor who provided helpful and insightful comments.

Data availability statement. Bias correction and quality control (BCQC) SNOTEL data are available at https://www.pnnl.gov/data-products, as developed and referenced in Yan et al. (2018). Additional hourly temperature data from SNOTEL sites were downloaded from the Natural Resources Conservation Service (NRCS) portal at https://wcc.sc.egov.usda.gov/reportGenerator/. Analysis of Record for Calibration (AORC) 1-km gridded forcings are not publicly available

but were provided to the project team by the Office of Water Prediction (OWP) of the National Oceanic and Atmospheric Administration's National Weather Service. The WRF preprocessing system (WPS) codes are available at https://github.com/wrf-model/WPS. The latest Noah-MP community model repository is available at https://github.com/NCAR/noahmp and WRF-Hydro at https://github.com/NCAR/wrf_hydro_nwm_public. Model outputs developed in the study are available from the authors upon request.

REFERENCES

Abolafia-Rosenzweig, R., C. He, S. P. Burns, and F. Chen, 2021: Implementation and evaluation of a unified turbulence parameterization throughout the canopy and roughness sublayer in Noah-MP snow simulations. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002665, https://doi.org/10.1029/2021MS002665.

——, ——, S. McKenzie Skiles, F. Chen, and D. Gochis, 2022: Evaluation and optimization of snow albedo scheme in Noah-MP land surface model using in situ spectral observations in the Colorado Rockies. *J. Adv. Model. Earth Syst.*, **14**, e2022MS003141, https://doi.org/10.1029/2022MS003141.

Barlage, M., and Coauthors, 2010: Noah land surface model modifications to improve snowpack prediction in the Colorado Rocky Mountains. *J. Geophys. Res.*, **115**, D22101, https://doi.org/10.1029/2009JD013470.

Chen, F., Z. Janjić, and K. Mitchell, 1997: Impact of atmospheric surface-layer parameterizations in the new land-surface scheme of the NCEP mesoscale eta model. *Bound.-Layer Meteor.*, **85**, 391–421, https://doi.org/10.1023/A:1000531001463.

Cho, E., C. M. Vuyovich, S. V. Kumar, M. L. Wrzesien, R. S. Kim, and J. M. Jacobs, 2022: Precipitation biases and snow physics limitations drive the uncertainties in macroscale modeled snow water equivalent. *Hydrol. Earth Syst. Sci.*, **26**, 5721–5735, https://doi.org/10.5194/hess-26-5721-2022.

Cosgrove, B., and Coauthors, 2024: NOAA's National Water Model: Advancing operational hydrology through continental-scale modeling. *J. Amer. Water Resour. Assoc.*, **60**, 247–272, https://doi.org/10.1111/1752-1688.13184.

Cosgrove, B. A., and Coauthors, 2003: Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *J. Geophys. Res.*, **108**, 8842, https://doi.org/10.1029/2002JD003118.

Dickinson, R. E., A. Henderson-Sellers, P. J. Kennedy, and M. F. Wilson, 1986: Biosphere-atmosphere Transfer Scheme (BATS) for the NCAR Community Climate Model. NCAR Tech. Note NCAR/TN-275+STR, 82 pp., https://doi.org/10.5065/D6668B58.

Dozier, J., E. H. Bair, and R. E. Davis, 2016: Estimating the spatial distribution of snow water equivalent in the world's mountains. *Wiley Interdiscip. Rev.: Water*, **3**, 461–474, https://doi.org/10.1002/wat2.1140.

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

Gochis, D. J., and Coauthors, 2018: The WRF-Hydro modeling system technical description, (version 5.0). NCAR Tech. Note, 107 pp., https://ral.ucar.edu/sites/default/files/public/WRF-HydroV5TechnicalDescription.pdf.

Guan, B., N. P. Molotch, D. E. Waliser, E. J. Fetzer, and P. J. Neiman, 2010: Extreme snowfall events linked to atmospheric rivers and surface air temperature via satellite measurements. *Geophys. Res. Lett.*, **37**, L20401, https://doi.org/10.1029/2010GL044696.

He, C., F. Chen, R. Abolafia-Rosenzweig, K. Ikeda, C. Liu, and R. Rasmussen, 2021: What causes the unobserved early-spring snowpack ablation in convection-permitting WRF modeling over Utah Mountains? *J. Geophys. Res. Atmos.*, **126**, e2021JD035284, https://doi.org/10.1029/2021JD035284.

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

——, M. Russo, and M. Anderson, 2016: Predictability of seasonal streamflow in a changing climate in the Sierra Nevada. *Climate*, **4**, 57, https://doi.org/10.3390/cli4040057.

Jordan, R., 1991: A one-dimensional temperature model for a snow cover: Technical documentation for SNTHERM.89. 62 pp., https://erdc-library.erdc.dren.mil/jspui/bitstream/11681/11677/1/SR-91-16.pdf.

Kim, R. S., and Coauthors, 2021: Snow Ensemble Uncertainty Project (SEUP): Quantification of snow water equivalent uncertainty across North America via ensemble land surface modeling. *Cryosphere*, **15**, 771–791, https://doi.org/10.5194/tc-15-771-2021.

Kitzmiller, D. H., W. Wu, Z. Zhang, N. Patrick, and X. Tan, 2018: The analysis of record for calibration: A high-resolution precipitation and surface weather dataset for the United States. *2018 Fall Meeting*, Washington, D.C., Amer. Geophys. Union, Abstract H41H-06, https://ui.adsabs.harvard.edu/abs/2018AGUFM.H41H..06K/abstract.

Letcher, T. W., J. R. Minder, and P. Naple, 2022: Understanding and improving snow processes in Noah-MP over the northeast United States via the New York state Mesonet. ERDC Tech. Rep. ERDC/CRREL TR-22-9, 50 pp., https://doi.org/10.21079/11681/45060.

Lettenmaier, D. P., D. Alsdorf, J. Dozier, G. J. Huffman, M. Pan, and E. F. Wood, 2015: Inroads of remote sensing into hydrologic science during the WRR era. *Water Resour. Res.*, **51**, 7309–7342, https://doi.org/10.1002/2015WR017616.

Li, J., C. Miao, G. Zhang, Y.-H. Fang, W. Shangguan, and G.-Y. Niu, 2022: Global evaluation of the Noah-MP land surface model and suggestions for selecting parameterization schemes. *J. Geophys. Res. Atmos.*, **127**, e2021JD035753, https://doi.org/10.1029/2021JD035753.

Li, Q., T. Yang, and L. Li, 2022: Quantitative assessment of the parameterization sensitivity of the WRF/Noah-MP model of snow dynamics in the Tianshan Mountains, central Asia. *Atmos. Res.*, **277**, 106310, https://doi.org/10.1016/j.atmosres.2022.106310.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res.*, **99**, 14 415–14 428, https://doi.org/10.1029/94JD00483.

Liston, G. E., and K. Elder, 2006: A meteorological distribution system for high-resolution terrestrial modeling (MicroMet). *J. Hydrometeor.*, **7**, 217–234, https://doi.org/10.1175/JHM486.1.

Niu, G.-Y., and Z.-L. Yang, 2007: An observation-based formulation of snow cover fraction and its evaluation over large North American river basins. *J. Geophys. Res.*, **112**, D21101, https://doi.org/10.1029/2007JD008674.

——, and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements.

J. Geophys. Res., **116**, D12109, https://doi.org/10.1029/2010JD015139.

Painter, T. H., A. P. Barrett, C. C. Landry, J. C. Neff, M. P. Cassidy, C. R. Lawrence, K. E. McBride, and G. L. Farmer, 2007: Impact of disturbed desert soils on duration of mountain snow cover. *Geophys. Res. Lett.*, **34**, L12502, https://doi.org/10.1029/2007GL030284.

——, S. M. Skiles, J. S. Deems, A. C. Bryant, and C. C. Landry, 2012: Dust radiative forcing in snow of the Upper Colorado River Basin: 1. A 6 year record of energy balance, radiation, and dust concentrations. *Water Resour. Res.*, **48**, W07521, https://doi.org/10.1029/2012WR011985.

Pan, M., and Coauthors, 2003: Snow process modeling in the North American Land Data Assimilation System (NLDAS): 2. Evaluation of model simulated snow water equivalent. *J. Geophys. Res.*, **108**, 8850, https://doi.org/10.1029/2003JD003994.

Sakaguchi, K., and X. Zeng, 2009: Effects of soil wetness, plant litter, and under-canopy atmospheric stability on ground evaporation in the Community Land Model (CLM3.5). *J. Geophys. Res.*, **114**, D01107, https://doi.org/10.1029/2008JD010834.

Serreze, M. C., M. P. Clark, and A. Frei, 2001: Characteristics of large snowfall events in the montane western United States as examined using snowpack telemetry (SNOTEL) data. *Water Resour. Res.*, **37**, 675–688, https://doi.org/10.1029/2000WR900307.

Sun, N., H. Yan, M. S. Wigmosta, L. R. Leung, R. Skaggs, and Z. Hou, 2019: Regional snow parameters estimation for large-domain hydrological applications in the western United States. *J. Geophys. Res. Atmos.*, **124**, 5296–5313, https://doi.org/10.1029/2018JD030140.

——, ——, ——, J. Lundquist, S. Dickerson-Lange, and T. Zhou, 2022: Forest canopy density effects on snowpack across the climate gradients of the western United States mountain ranges. *Water Resour. Res.*, **58**, e2020WR029194, https://doi.org/10.1029/2020WR029194.

Tanaka, S. K., and Coauthors, 2006: Climate warming and water management adaptation for California. *Climatic Change*, **76**, 361–387, https://doi.org/10.1007/s10584-006-9079-5.

Vicuña, S., R. D. Garreaud, and J. McPhee, 2011: Climate change impacts on the hydrology of a snowmelt driven basin in semiarid Chile. *Climatic Change*, **105**, 469–488, https://doi.org/10.1007/s10584-010-9888-4.

Wang, Y.-H., P. Broxton, Y. Fang, A. Behrangi, M. Barlage, X. Zeng, and G.-Y. Niu, 2019: A wet-bulb temperature-based rain-snow partitioning scheme improves snowpack prediction over the drier western United States. *Geophys. Res. Lett.*, **46**, 13 825–13 835, https://doi.org/10.1029/2019GL085722.

Wiken, E., F. Jiménez Nava, and G. Griffith, 2011: *North American Terrestrial Ecoregions—Level III*. Commission for Environmental Co-operation, 149 pp.

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

You, Y., C. Huang, Z. Yang, Y. Zhang, Y. Bai, and J. Gu, 2020: Assessing Noah-MP parameterization sensitivity and uncertainty interval across snow climates. *J. Geophys. Res. Atmos.*, **125**, e2019JD030417, https://doi.org/10.1029/2019JD030417.

Zhang, G., G. Zhou, F. Chen, M. Barlage, and L. Xue, 2014: A trial to improve surface heat exchange simulation through sensitivity experiments over a desert steppe site. *J. Hydrometeor.*, **15**, 664–684, https://doi.org/10.1175/JHM-D-13-0113.1.

——, F. Chen, and Y. Gan, 2016: Assessing uncertainties in the Noah-MP ensemble simulations of a cropland site during the Tibet Joint international cooperation program field campaign. *J. Geophys. Res. Atmos.*, **121**, 9576–9596, https://doi.org/10.1002/2016JD024928.