Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR026853

Special Section:
Advances in remote sensing, measurement, and simulation of seasonal snow

Improving SWE Estimation With Data Assimilation: The Influence of Snow Depth Observation Timing and Uncertainty

Eric J. Smyth[1], Mark S. Raleigh[1,2,3], and Eric E. Small[1]

[1]Department of Geological Sciences, University of Colorado, Boulder, CO, USA, [2]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, CO, USA, [3]National Snow and Ice Data Center (NSIDC), University of Colorado, Boulder, CO, USA

Key Points:
- Data assimilation improves modeled snow depth, density, and SWE across a range of climates and wet/dry seasons in the western United States
- Assimilating one snow depth observation near maximum accumulation yields the majority of improvements to estimated snow density and SWE
- Assimilating frequent snow depth observations yields the lowest SWE uncertainty and smallest melt-out date errors

Supporting Information:
- Supporting Information S1

Correspondence to:
E. J. Smyth,
eric.smyth@colorado.edu

Citation:
Smyth, E. J., Raleigh, M. S., & Small, E. E. (2020). Improving SWE estimation with data assimilation: The influence of snow depth observation timing and uncertainty. *Water Resources Research*, *56*, e2019WR026853. https://doi.org/10.1029/2019WR026853

Received 27 NOV 2019
Accepted 13 APR 2020
Accepted article online 16 APR 2020

Abstract Snow depth observations can be leveraged with data assimilation (DA) to improve estimation of snow density and snow water equivalent (SWE). A key consideration for mission and campaign design is how snow depth retrieval characteristics (including observation timing/frequency and sampling error) influence SWE accuracy and uncertainty in a DA framework. To quantify these effects, we implement a particle filter (PF) assimilation technique to assimilate depth and validate this approach against observed snow density and SWE at 49 snow telemetry sites across 9 years. We sample from continuous in situ snow depth records to test a range of measurement timing and sampling error scenarios representative of remote sensing capabilities. Assimilation reduces density bias by over 40% and SWE bias by over 70% across climate zones and in both wet and dry years. There is little incremental benefit to SWE accuracy when assimilating more than one depth observation near peak accumulation. SWE estimates are less sensitive to observation timing than sampling error. Alternatively, more frequent depth observations improve melt-out date timing and reduce SWE uncertainty, a key consideration when evaluating the operational utility of DA. In matching depth observations, the PF mostly acts to increase model precipitation inputs, while not systematically shifting other parameter values or forcings across the climate zones represented with the study sites. This demonstrates that precipitation is the largest source of model error. With DA, density errors are still nontrivial (above 10%), illuminating the need for further improvements to modeled density to estimate SWE within specified error limits.

Plain Language Summary The amount of water stored in seasonal snowpack (snow water equivalent or "SWE") is an important variable for water management yet is currently difficult to measure in mountainous areas. One technique is to measure snow depth from airborne or satellite platforms and use that depth to guide a model that simulates snow density and SWE with a technique called data assimilation. We show that assimilation reduces errors in modeled SWE relative to control model runs, even when using a single depth observation to guide the model. However, more depth observations are helpful to reduce model uncertainty.

1. Introduction

Mountain snowpack and its melt dominate the surface hydrology of many regions, with implications for water supply, hydropower, ecological processes, weather, and regional and global climate (Bales et al., 2006; Serreze et al., 1999). Ongoing campaigns (e.g., NASA SnowEx) aim to determine the optimal combination of remote sensors, platforms, and models needed to map snow water equivalent (SWE) globally—which is not currently possible from existing satellite platforms (Dozier, 2011; Nolin, 2010; Sturm, 2015). The recent Decadal Survey (National Academies of Sciences, Engineering, and Medicine, 2018) has recommended an "explorer priority" to global snow depth and SWE mapping, with capabilities for resolving snow amounts in mountain areas. These regional and global missions, as well as more local measurement campaigns, prompt questions about the optimal characteristics of snow mapping with remote sensing and modeling frameworks.

Spatially extensive snow depth observations can be used as the basis for maps of SWE (e.g., Painter et al., 2016). Snow depth mapping is possible with several remote sensing approaches, including lidar, radar, and photogrammetry, with deployments possible on spaceborne, airborne, drone, and ground-based

platforms (e.g., Grünewald et al., 2010; Lievens et al., 2019; Marti et al., 2016; Moller et al., 2017; Painter et al., 2016; Vander Jagt et al., 2015). Depending on the sensor, platform, and spatial scale, repeat altimetry snow depth observations have errors of ~10–50 cm (Deems et al., 2013; Lievens et al., 2019; Marti et al., 2016). Depth mapping efforts have been completed at plot to watershed scales, but spatially extensive depth mapping is becoming possible at larger scales (e.g., mountain ranges). Retrieval methods over these larger domains have a wide range of temporal sampling frequency; satellite-based mapping may have an infrequent and fixed revisit time (e.g., potential snow depth retrievals every 91 days at mid-latitudes with ICESat-2), while airborne approaches (e.g., Airborne Snow Observatory, ASO) are more flexible and can target key periods in the seasonal snow cycle (e.g., peak accumulation and melt season). While some current/future remote sensing techniques aim to measure SWE directly (e.g., passive microwave, Global Navigation Satellite System, or radar-based techniques), SWE mapping based on spatially extensive snow depth observations provide higher spatial resolution and likely smaller errors (e.g., Painter et al., 2016).

Snow depth alone is not sufficient to map SWE—estimates of snow bulk density are required to convert snow depth into an equivalent water depth (e.g., Painter et al., 2016). However, there has been no demonstrated capability for mapping snow density with satellite or airborne platforms, and in situ density observations are too sparse to be combined with spatially extensive depth observations. Thus, model-based estimates of bulk snow density are required to calculate SWE from intermittent repeat depth observations. These models also provide spatial estimates of SWE before, between, and after the times when snow depth is mapped. When snow depth distributions have been mapped across mountainous basins, SWE uncertainty is dominated by uncertainty in modeled density (Raleigh & Small, 2017). Uncertainty in modeled density results from forcing data (e.g., high uncertainty in precipitation), parameterizations (e.g., physical snow compaction routines), and model choice of process representations. For example, estimated density can vary over a wide range (e.g., $\pm$30%) due to model selection alone (Feng et al., 2008). Thus, refining SWE estimates derived from repeat mapping of snow depth depends on improving modeled bulk density, including a reduction in both errors (e.g., root mean squared error, RMSE) and uncertainty (i.e., standard deviation).

Data assimilation (DA) has been used to combine observations of snow depth with the aforementioned imperfect model simulations of snow bulk density. As expected, simulated snow depth is improved when observations of snow depth are assimilated, a result consistent across DA techniques (Hedrick et al., 2018; Magnusson et al., 2017; Margulis et al., 2019). Assimilation of depth also provides a mechanism to improve modeled bulk density: When simulated depth is too low (high) due to a low (high) precipitation bias, simulated density can be too low (high) because of insufficient overburden compaction over time. Smyth et al. (2019) demonstrated how assimilation of intermittent depth observations can reduce model bulk density errors by 25–50%, leading to error reductions in SWE relative to a control model. While the application of DA to intermittent snow depth observations is nascent, this approach can indirectly improve snow density to further improve SWE estimates, improvements that are not produced when combining an independent model estimate of density with observed snow depth.

The design of a remote sensing mission to map SWE from snow depth is necessarily focused on how characteristics of the sensor and platform affect observations of depth. But it must also consider how these characteristics interact with models of snow bulk density and the assimilation system that is used to estimate SWE from depth observations. Regardless of the spatial scale of the snow depth observations, there are two aspects of depth retrievals that have the potential to influence DA: observation uncertainty and sampling frequency. Sampling frequency determines the number of assimilation adjustments to modeled states, and thus potentially the accuracy of the system. Each approach to satellite-based snow depth mapping has a different revisit time, determined by sensor and satellite characteristics (e.g., NASA's ICESat-2 lidar and ESA's Sentinel-1 SAR). Airborne platforms (e.g. ASO lidar and drone-based photogrammetry) can be deployed to target specific times (e.g., peak SWE). Thus, there is a wide range of possible revisit times or sampling dates within a snow season. The magnitude of observational uncertainty directly influences the outcome of DA: Larger uncertainty results in less pronounced adjustments to modeled states (van Leeuwen, 2009). In addition, every combination of sensor and platform yields a different magnitude of snow depth uncertainty: Airborne lidar has an uncertainty of approximately 10 cm (e.g., Deems et al., 2013), space-based SAR uncertainty is approximately 30 cm (Lievens et al., 2019), and satellite-based

photogrammetry uncertainty can exceed 50 cm (e.g., Marti et al., 2016). In this paper, we implement a framework to assimilate snow depth observations to investigate the effects of both observation timing and sampling error on model estimates of SWE.

2. Objectives and Approach

Previous snow assimilation studies have investigated the impact of observation timing on model simulations. In two studies, the greatest gains in season-wide model depth accuracy were achieved by assimilating a single depth survey, with inclusion of additional surveys yielding smaller, incremental benefits (Hedrick et al., 2018; Margulis et al., 2019). These studies focused only on improvements to estimated snow depth in a single basin in California and were temporally limited to assimilating airborne lidar data available from ASO (weekly–biweekly from March–summer). Garnaud et al. (2019) assimilated SWE and found a shorter revisit time (i.e., more observations) to be the key parameter for improving model estimates. However, rather than focusing on a state variable (SWE), this study focused on simulating a flux (runoff), which responds to individual precipitation events which were important to capture with frequent observations. Therefore, prior research has not established the utility of more frequent observations for the purpose of improving modeled density and therefore SWE. Addressing this research gap holds implications for basic research and science applications in watershed forecasting and management. For example, California Senate Bill 487 (introduced 21 February 2019) would have required the state to conduct up to 10 airborne snow depth surveys per year over major reservoir-supplying basins, implying an added value to observing snow depth many times throughout the water year (WY). These and other spatially extensive, repeat snow depth retrievals can be costly (e.g., $150 million over 10 years in CA Bill 487, Governor's Veto Message 13 October 2019), and thus it is critical to assess the gains offered by additional snow depth observations from a remote sensing mission or field campaign.

Of similar importance is the precision of snow depth observations. The Direct Insertion methodology employed by Hedrick et al. (2018) treats snow depth observations as error free. Margulis et al. (2019) assumes 5 cm sampling error in the analysis, appropriate for airborne lidar retrievals. The study references a sensitivity test with sampling error ranging between 2 and 5 cm that produced limited differences in results. It is not well established how estimated density and SWE in a DA framework respond to larger errors in measured snow depth (e.g., in excess of 5 cm), errors which are possible with multiple platforms/techniques.

Here, we explore several factors critical to mission design that have not been considered in these studies above. First, we quantify DA improvements to modeled snow density, SWE, and melt-out dates, not just snow depth. Second, we consider not just the accuracy of SWE estimates but also their uncertainty, which is critical in operational applications. Third, we evaluate multiple snow depth sampling error magnitudes to reflect a wider range of depth sensor precisions, not just airborne lidar. A key consideration for designing regional and global SWE missions (e.g., NASA SnowEx and Decadal Survey) is the robustness of a snow measurement-modeling strategy across regional and interannual variations in climate. To examine these dependencies, we conduct experiments at sites across the major snow regions of the western United States across a 9-year interval, WYs 2008–2016, to examine a wide range of depth and density conditions that exist (Bormann et al., 2013; Sturm et al., 2010; Sturm & Holmgren, 1998).

While different quantitative goals of SWE error and uncertainty are appropriate for different spatial scales and applications, the Decadal Survey Water Resource Panel targets SWE error of 10% (National Academies of Sciences, Engineering, and Medicine, 2018). We will refer to this benchmark and are interested in which characteristics of snow depth acquisition (measurement frequency and uncertainty) allow our DA system to achieve this benchmark (or which types come closest to meeting this benchmark, if none can achieve it). In some circumstances, performance could be improved beyond that achieved by assimilation alone, for example, by employing model calibration or bias adjustment. We show below that precipitation inputs are the largest source of model error and that the DA system does not consistently shift the most sensitive model density parameter. Thus, while calibration can improve model density in some sites and years, it is not the most effective way to improve SWE estimates over wide domains and in different environments.

We first demonstrate the utility of DA by improving model estimates relative to a set of control (or open loop, OL) model runs (section 4.1). This comparison can have limited utility, because OL simulations can be

deceptively poor without model calibration. Therefore, we focus on intercomparing assimilation results in different snow depth observation scenarios (Section 4.2), to isolate the sensitivity of SWE error and uncertainty to observation timing and sampling error. Through these latter comparisons, we compare different timing/error scenarios to the 10% SWE error benchmark articulated in the Decadal Survey.

3. Data, Methods, and Experiments

3.1. Snow Data

The goal of assimilating snow depth observations is to improve estimates of SWE. Central to this effort is a coincident reduction in snow density errors. Therefore, measuring performance of a snow DA system requires collocated snow depth, density, and SWE data. While assimilation of remotely sensed, spatially extensive snow depth data is the end goal, these data are not accompanied by sufficient observations of snow density and SWE, measured over the same scale of spatial support as snow depth. This does not permit effective validation of DA estimates of SWE or density, a limitation recognized by prior studies (e.g., Margulis et al., 2019).

Here, we sample the continuously observed snow depth time series recorded at snow telemetry (SNOTEL) sites to generate synthetic records of intermittent snow depth observations. These records serve as a proxy for remotely sensed observations (e.g., differential altimetry from lidar or SAR), with the benefit of collocated depth, density, and SWE data that are continuous, allowing for evaluation of errors both at the time of assimilation as well as before and after assimilation steps. Figure 1a demonstrates a continuous depth data set resampled to monthly intermittent observations. We detail the sampling scenarios in section 3.6.

We manually screened 930 sites in the contiguous United States for continuous snow depth and SWE records between WYs 2008 and 2016 and for locations with site photos that show depth sensors located directly above the snow pillows. This step is necessary because we validate estimated density at each site against density observations that we derive by dividing the snow pillow SWE by collocated ultrasonic snow depth and then normalizing by liquid water density (e.g., Mizukami & Perica, 2008). From this subset of sites, we selected 49 sites across the western United States (Figure 2a) to represent a range in proximity to large bodies of water, elevation, wintertime temperature, and precipitation—all factors known to influence snow depth and density regimes (e.g., Mizukami & Perica, 2008). We chose at least one site from each of the eight climate zones identified in Serreze et al. (1999), an additional 10 maritime sites on the western slope of the Cascades (Zone 1: Pacific Northwest), 10 intermountain sites on the eastern slope of the Sierra Nevada (Zone 2: Sierras), 10 intermountain sites on the eastern edge of the basin-and-range province (Zone 6: Utah), and 10 continental sites within the Colorado Rockies (Zone 7: Colorado). Figure 2a plots the site locations, and Table S1 in the supporting information provides summary information. Melt-out dates were recorded at each site as the first snow-free date (i.e., SWE < 25 mm) after the date of peak SWE. We did not include sites from the California Department of Water Resources (CDWR) network (as in Margulis et al., 2019), as information regarding collocation of CDWR snow pillows and depth sensors was not readily available to us.

We utilize the SNOTEL network to assess the performance of the particle filter (PF) in different climates and with different revisit/sampling intervals but recognize disadvantages to this approach. First, SNOTEL sites are preferentially located in flat, open clearings at midelevation—and can be unrepresentative of local conditions (e.g., Gleason et al., 2017; Molotch & Bales, 2005; Wetlaufer et al., 2016). Second, snow bridging over pillows can cause undersampling of SWE, and therefore density derived from SWE (Serreze et al., 1999). Finally, we use SNOTEL data for both assimilation (depth) and validation (density and SWE). Although depth and SWE are measured independently, the density observations used for validation are derived in part from the depth we assimilate, which is not ideal. Still, these sites provide the density and SWE data required for evaluation of the PF technique that would otherwise not be available at a location with snow depth observations alone (e.g., at all grid cells in a typical mountain basin).

3.2. Model Forcing Data

Most SNOTEL sites do not record the suite of inputs required for process-based snow models (Raleigh et al., 2016). Therefore, we force our model with hourly 1/8-degree data from the NASA North American Land Data Assimilation System Phase 2 (NLDAS-2, hereafter called NLDAS) grid cells that contain each site (Xia et al., 2012). This type of gridded product could be used to generate operational SWE products across

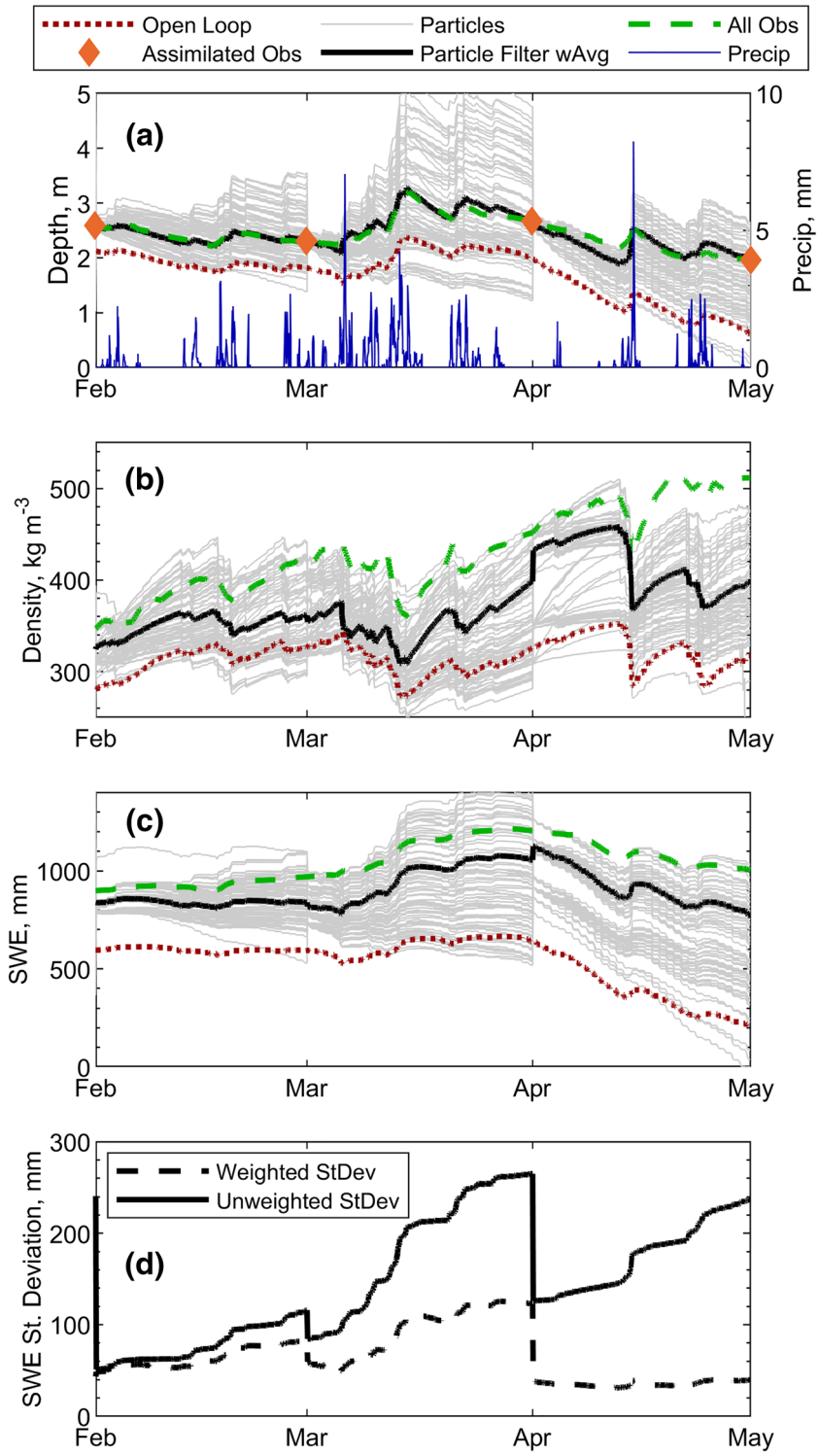

Figure 1. Example implementation of PF improvements to (a) snow depth, (b) density, and (c) snow water equivalent (SWE). In panel (a), continuous depth observations (green dashed line) are resampled to monthly intermittent observations (orange diamonds). The particle filter "wAvg" is the weighted average of particle values, with weights generated by the PF. Precipitation values are daily sums. Panel (d) shows uncertainty, with lines for both weighted and unweighted standard deviation.

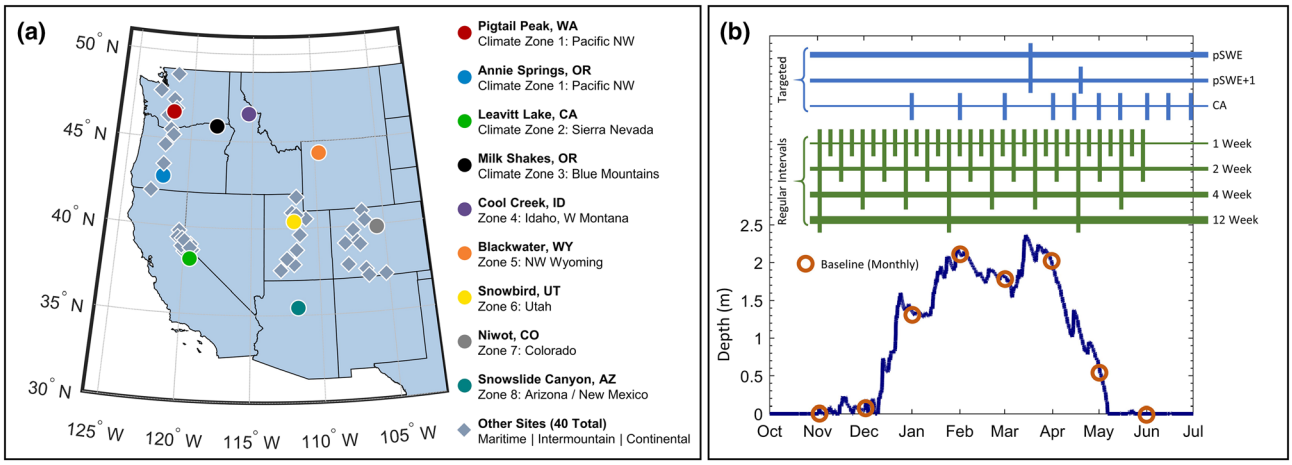

Figure 2. Panel (a) shows a map of SNOTEL locations. Colored circle locations are from regions identified in Serreze et al. (1999). Other sites (diamonds) span continental to maritime climates. Panel (b) is a visualization of snow depth sampling scenarios at a given site, relative to an example depth time series. For each timing scenario (represented by a horizontal bar), depth would be sampled at each vertical tick. Here, pSWE and pSWE+1 are shown relative to this example's date of peak SWE, whereas we actually calculate a climatological date of peak SWE with historical data.

drainage basins and is therefore appropriate for this study. Precipitation and temperature are measured at all SNOTEL sites. However, we did not combine these on-site observations with NLDAS forcing data as a blended input data set would not be internally consistent. More specifically, this blending could result in nonrealistic combinations of forcing data, such as a day where SNOTEL data suggest precipitation is occurring but NLDAS radiation data suggest clear-sky conditions.

In generating our main results, we do not downscale NLDAS forcing inputs for two reasons. First, Smyth et al. (2019) showed that density and SWE errors with assimilation and nondownscaled NLDAS model forcing were similar to errors when using high-quality local meteorological forcing. Second, our DA accounts for meteorological uncertainty and thus has some capability in adjusting for systematic biases in forcing data (a function similar to downscaling). Our sites are flat, so downscaling would largely involve adjusting precipitation, temperature, etc. based on the elevation difference between site and NLDAS grid cells and lapse rates (e.g., from PRISM). This elevation difference averages ~200 m across our sites, so average changes to precipitation (~5%) and temperature (~1.5 °C) with downscaling would already be within the range of variations we generate in the DA scheme (see Text S3) and are therefore unnecessary. Similarly, shortwave and longwave radiation vary with elevation and are meteorological inputs we vary in the DA framework (Text S3).

To support our decision not to downscale in the main analysis, we perform a test that demonstrates that downscaling precipitation and temperature inputs prior to assimilation yields negligible differences in SWE, as predicted given the small elevation differences between SNOTEL stations and NLDAS cells. Even with downscaling, significant changes to precipitation inputs are required for assimilation to match depth observations (Figure S1). We also compare NLDAS precipitation to gauge measurements at each SNOTEL site. Precipitation at SNOTEL sites is 30% higher (on average) than from NLDAS (Figure S2), a difference similar to values reported previously (He et al., 2019).

3.3. Snow Model

For this study, we use the Flexible Snow Model (FSM2), an extension of the Factorial Snow Model (Essery, 2015). FSM2 (version released 15 November 2018) is a multiphysics energy balance model of snow accumulation and melt—with modular options for seven parameterizations, giving 192 possible model configurations. Model physics and configuration options are documented at https://github.com/RichardEssery/FSM2.

We compiled FSM2 with the default physics options. When fresh snow is deposited, the snowpack mass increases by the amount of precipitation, new density is fixed to 100 kg/m^3, and the snowpack depth increases by the ratio of the two. Density then changes as a function of viscous overburden compaction by

overlying layers and thermal metamorphism, using equations developed by Anderson (1976). SWE additionally changes with modeled melt and sublimation. Our sites are flat and open, so we do not use FSM2 to simulate canopy dynamics. We use the default option for the number of snow layers ($n = 3$), with maximum thickness of 10 cm in the top layer, 20 cm in the middle layer, and no maximum thickness in the bottom layer. We do not experiment with varying the number or thickness of snow layers, but discuss it as a possible avenue for improving modeled density (section 5).

3.4. DA Methodology

The DA technique used is the PF (Smyth et al., 2019). The PF attempts to represent the true probability density of a state variable by generating an ensemble of model estimates or particles (Figures 1a to 1c). The particles diverge because the model runs are varied to reflect uncertainties in parameter choices and forcing data (section 3.5). The PF then weights each particle by comparing its state to available snow depth observations (Figure 1a), while also accounting for the observational uncertainty. To avoid filter degeneracy (van Leeuwen, 2009), the particles are resampled after each assimilation timestep—wherein particles with very low weights are eliminated, while particles with high weights are duplicated and advanced to the next assimilation timestep. This is why the spread between particles contracts at assimilation points in Figures 1a to 1c. Weighting and resampling do not alter model states—similar to the Particle Batch Smoother but unlike direct insertion and Kalman update approaches.

We propagate 75 particles between assimilation timesteps following particle-number sensitivity data in Magnusson et al. (2017). We use the stochastic universal (re)sampling algorithm, which Kitagawa (1996) showed to have the lowest resampling noise among several methods (van Leeuwen, 2009). With this algorithm, the weights for each particle (which sum to 1) are placed on a line [0,1]. Then, a random number is chosen between 0 and 1/N, and then N segments of length 1/N are added in sequence. Particles are resampled if a line endpoint falls within its weight bin. In this way, some particles can be chosen more than once, but N particles are always selected for propagation.

We calculate the PF "best estimate" of snowpack depth, density, and SWE as the weighted average of all particle simulations over time, with weights that are generated by the PF at every assimilation timestep and carried backwards in time through the interval preceding the assimilated observation (Magnusson et al., 2017). We also generate an OL control run for comparison at all sites and years, using unperturbed NLDAS forcings and the default parameters in FSM2 (Figures 1a to 1c). This OL simulation is the equivalent of the simulation one would use in the absence of DA. To evaluate the PF and OL accuracy relative to observations, we calculate bias and RMSE in density and SWE and compare errors in estimated melt-out dates. To evaluate the uncertainty in SWE estimates, we calculate the weighted standard deviation (1-sigma) of particles through time (Figure 1d) with weights generated by the PF. Uncertainty is an important consideration when evaluating the operational utility of the PF, especially between observations when particles have not been reweighted and the particle spread and uncertainty increases (Figures 1a and 1d). By this metric, uncertainty is similar to precision and is governed by the spread of particles and their weights—which in turn are influenced by the timing and uncertainty of observations (Figure S3).

3.5. PF Ensemble Generation

Following procedure from Smyth et al. (2019), we generate ensembles of particles by running the model with different forcing inputs and varying a snow compaction parameter—to represent possible sources of model uncertainty (Table 1).

We vary the snow viscosity parameter (called "etab" in FSM2 code) around the commonly used default value of 21 $cm^3\ g^{-1}$ (Kojima, 1967) in a uniform range ($\pm 6\ cm^3\ g^{-1}$) to represent uncertainty in model parameter values. This single parameter was selected based on a parameter sensitivity analysis, which showed variations in etab accounted for almost all of variations in SWE when all densification parameters were varied and no forcing data uncertainty was considered (Figure S4). In preliminary experiments, we found that varying new snow density ("rhof") did not influence results, so we exclude it here for simplicity.

We perturb each particle's hourly precipitation, radiation, and temperature inputs with additive stochastic noise, pulled from normal distributions with bounds given in Table 1. To capture systematic errors, we also apply a systematic bias to each forcing—a single bias perturbation applied to an entire "window" between subsequent assimilation timesteps, following ranges applied in Raleigh et al. (2015). This two-step

Table 1
Details on Perturbations to a Snow Viscosity Parameter (etab) and Various Forcing Data to Create the PF Ensemble

Variable	Unit	Adjustment	Distribution	Lower bound	Upper bound
etab	cm^3/g	Additive	Uniform	−6	+6
Precipitation					
Hourly noise	mm/h	Additive	Normal	−25%	+25%
Window adj.	mm/h	Multiplicative	Uniform	−75%	+300%
SW Radiation					
Hourly noise	W/m^2	Additive	Normal	−160	+160
Window adj.	W/m^2	Additive	Normal	−100	+100
LW Radiation					
Hourly noise	W/m^2	Additive	Normal	−80	+80
Window adj.	W/m^2	Additive	Normal	−25	+25
Temperature					
Hourly noise	C	Additive	Normal	−7.5	+7.5
Window adj.	C	Additive	Normal	−3	+3

Note. Hourly noise adjusts each hour of input data individually, while the "window adjustment" refers to a bias (additive or multiplicative) applied to an entire time period between two assimilation steps.

approach to generate a forcing ensemble is similar to Magnusson et al. (2017). For example, precipitation is adjusted with an additive, hourly, random error pulled from a normal distribution bound between ±25% of the observed hourly data and a multiplicative window-wide bias pulled from a uniform distribution between −75% and 300%. All additive perturbations are positive, that is, the perturbed value is the observed value + perturbation. To bind each set of normally distributed perturbations in a desired range (e.g., ±80 W/m^3 for hourly shortwave radiation), we pull 75 (i.e., one for each particle) values from a standard Gaussian distribution (mean of zero, standard deviation of one) and then rescale and translate the set of numbers to have desired maximum and minimum values.

3.6. Experiments

First, we consider the performance of the PF relative to the OL across all sites and years, using a monthly observation interval and a single depth uncertainty of 10 cm (Figure 2b) to evaluate the capability of the PF technique across a range of snow climates and conditions.

Then, we investigate the effect of observation timing by evaluating PF performance in several possible retrieval scenarios:

A **Consistent revisit intervals** typical of existing space-based platforms but that could also be achieved with airborne missions under ideal weather conditions. These scenarios include 1-week, 2-week, 4-week, and 12-week intervals between observations—thereby covering a range of possible depth retrievals (e.g., near weekly from Sentinel-1 to approximately every 3 months from ICESat-2). For modeling purposes, we restrict observations to between November and June. A weekly interval yields 31 observations, a 2-week interval has 16 observations, a 4-week interval produces eight observations, and a 12-week interval has three observations (Figure 2b).

B **Targeted deployments** typical of airborne surveys. Specifically, we test scenarios with a single observation at peak SWE (pSWE), with observations at peak SWE and 1 month afterward (pSWE+1), and a series of 10 dates intended to represent the proposed sampling strategy over basins in California through Senate Bill 487 (we refer to as the "CA" scenario). The date of peak SWE for each site is calculated as the average day-of-year of peak SWE over the 9-year record (i.e., a climatological peak SWE date). The CA timing scenario calls for intensive observations during the late winter and melt season but does not specify the dates in Bill 487. Based on historic snow sampling strategies employed in California (snow courses and lidar surveys), we use a monthly sampling from January to April and biweekly thereafter through July (Figure 2b).

In all timing scenarios, we assume ideal conditions for observing snow depth, that is, no missing data or changes in observation error due to environmental factors known to impact some sensors, such as clouds, forests, or wet snow.

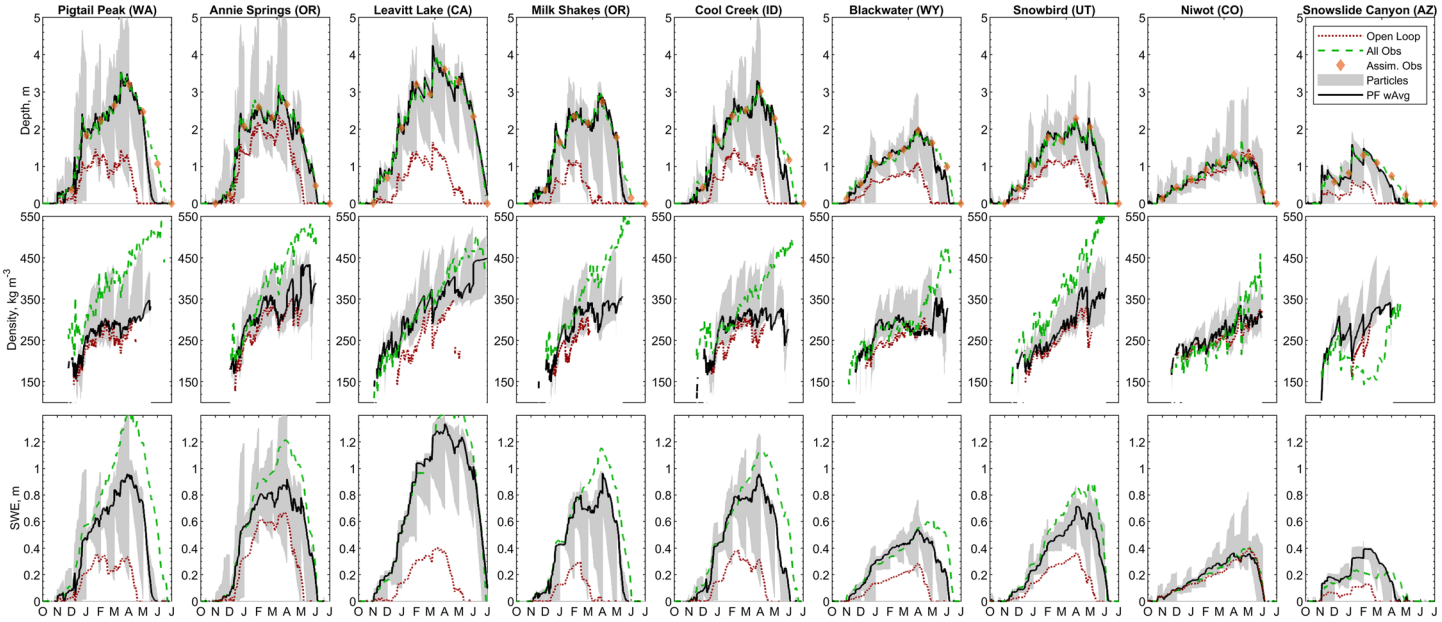

Figure 3. Comparison of observed data, OL, and PF estimates with assimilation of simulated monthly snow depth observations for nine selected SNOTEL sites—one in each zone from Serreze et al. (1999). All data are from WY 2016. At all sites, the PF is able to closely match observed depth. Improvements to density and SWE vary between sites (e.g., Leavitt Lake, CA vs. Snowslide Canyon, AZ).

Finally, we examine the impact of observational uncertainty by repeating our tests with different specified observation errors. We introduce a random error to the assimilated snow depth, pulled from a normal distribution with an RMSE ranging from 10 (e.g., airborne lidar) to 50 cm (e.g., spaceborne photogrammetry). We focus the results and discussion on these two end points (10 and 50 cm) in the analysis and then consider the full spectrum of observation error within this range together with the suite of timing scenarios to evaluate the linearity of results.

4. Results

4.1. Improvements From the PF

First, we compare OL and PF performance relative to in situ measurements at the 49 SNOTEL sites, simulating monthly depth observations for assimilation with observational uncertainty of 10 cm; sample results for WY 2016 are shown in Figure 3 for the nine sites we selected to span climate zones (see Figure 2a). These plots illustrate the PF framework at the point scale across wet or dry climates/years and in cases where the OL performs comparatively poorly or well relative to observations.

The spread between particles, and therefore the uncertainty of DA estimates, increases because we perturb model inputs and the densification parameter between assimilation steps. Particle spread tends to increase more during the accumulation season, because the model outputs are highly sensitive to precipitation, which is perturbed widely in our DA setup (Table 1, Figure S4). As a result, depth, density, and SWE standard deviation typically increases during the accumulation season and then level off or decrease at the end of the year, when precipitation is lower.

As expected, DA removes nearly all errors in snow depth (Figures 1a and 3). We focus our analysis instead on how DA improves simulations of snow density, SWE, and melt-out date. On average, density bias is reduced by 44%, SWE bias by 71%, and melt-out date by 87% (31 day improvement, on average) relative to the OL (Figure 4). The PF reduces SWE bias in 94% of cases, across sites and years ($n = 441$). Model accuracy improvement is similar with other metrics (e.g., RMSE, Table 2), at peak SWE (58% reduced SWE bias), and during the ablation season (57% reduced SWE bias). The OL nearly always underestimates density, SWE, and melt-out date; note the preponderance of positive OL errors in Figure 4, which signify model underestimation in the OL.

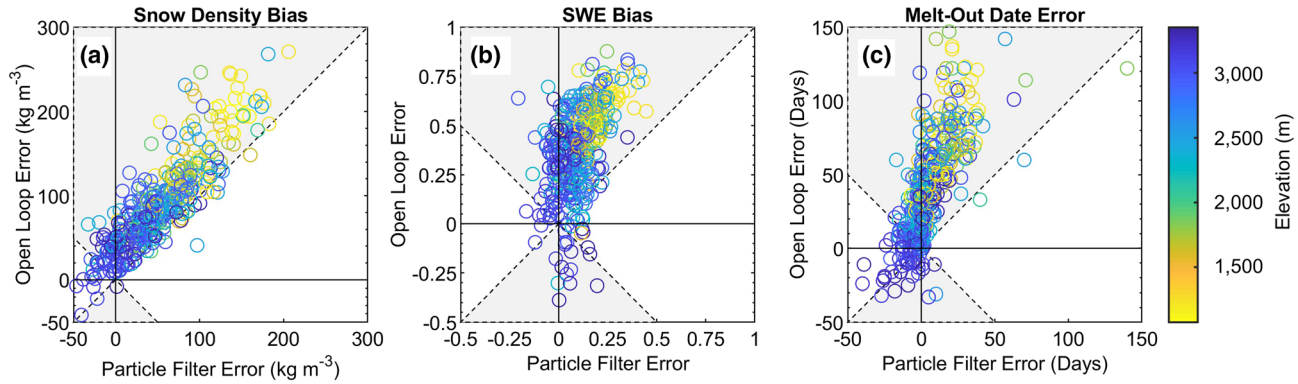

Figure 4. Scatterplots of (a) snow density bias, (b) SWE bias divided by peak SWE, and (c) melt-out date error for all sites and years, assuming 10 cm sampling error. Each point corresponds to an individual year at one of the 49 sites ($n = 441$) and each panel plots PF against OL performance with monthly observations, such that points in the gray areas indicate PF improvement. The sign convention is observation minus model, so positive values indicate the open loop model is underestimating density, SWE, and melt-out dates.

At all sites and years, depth assimilation improves model depth by weighting particles with considerably higher precipitation inputs than NLDAS through the snowfall correction factor (SCF, Figures 5a and 5b). The median SCF is 1.64, yielding precipitation input to the model that is over 50% greater than in the NLDAS forcing. The SCF is generally higher at warmer maritime sites (Figure 5a) and also higher at sites where SNOTEL elevation is greater than the elevation of the corresponding NLDAS grid cell (Figure 5b). It is possible that, in favoring particles with higher precipitation, the PF is also compensating for other model errors. However, the PF does not consistently shift other meteorological inputs (temperature or radiation) or the model parameter (snow viscosity) that we vary for ensemble generation (Figures 5c to 5f). One exception is with temperature in the maritime zone (Figure 5d), where the PF consistently favors particles with lower temperatures, which increases the snowfall fraction and reduces likelihood of midwinter melt events. Improved snow depth directly increases SWE estimates and thus extends estimated melt-out dates. It is likely that improved snow depth indirectly improves density estimates by increasing the amount of modeled overburden compaction.

4.2. Impact of Observation Timing and Errors on SWE Accuracy and Uncertainty

Having established the benefits of the PF relative to the OL across a wide range of annual and regional conditions in the case of monthly snow depth assimilation, we next examine how different snow depth sampling timing scenarios impact SWE accuracy (as reflected by RMSE) and uncertainty (as reflected by weighted standard deviation) and melt-out errors, focusing first on the 10 cm sampling error scenario. SWE accuracy has low sensitivity to the timing and total number of depth observations (Figure 6a). Across all years, we find only slight improvements in model accuracy with an increasing number of observations, either at regular intervals typical of satellites or campaign-style that can be applied in airborne or drone surveys. Only the pSWE and 12-week cases exhibit different performance than the rest (at the 5% significance level with a rank-sum statistical test, Table S2). However, both are only slightly worse than the other timing scenarios and all scenarios are far better than the OL (Table 3), which demonstrates that the majority of SWE RMSE improvements are attained by assimilating one depth observation (e.g., pSWE). Similar to the 10 cm error case, the PF performance with 50 cm sampling error is largely insensitive to the number of observations (Figure 6a).

The accuracy differences between timing scenarios are consistent across precipitation regimes and climates (Table 3). RMSE decreases moving from wetter, maritime climates (e.g., Pacific Northwest) to dryer, continental climates (e.g., Colorado). Additionally, we find the same patterns within each region as in the aggregate: More observations lead to lower model error, but all timing scenarios are significantly better than the OL.

SWE RMSE is sensitive to depth sampling error. For example, 50 cm sensor error yields ~30% greater RMSE than 10 cm sensor error (Table 2, Figure 6)—though it is still greatly improved over the OL. This difference in SWE RMSE between the two sampling error scenarios is significant when evaluating SWE RMSE against

Table 2
Comparison of Error Metrics Calculated Across all Sites and Years, With Monthly Sampled Observations

	RMSE		Bias			Melt-out date
Monthly observations	Full season	Abl. season	Full season	Abl. season	Peak SWE	Bias (days)
Density (kg/m^3)						
Open loop	93	146	80	142	101	40
PF: 10 cm sampling error	67	67	45	99	71	9
PF: 50 cm sampling error	69	69	47	103	75	14
SWE (mm)						
Open Loop	254	197	216	97	332	
PF: 10 cm Sampling Error	99	89	62	42	138	
PF: 50 cm Sampling Error	142	122	87	58	207	

Note. Ablation season is defined as peak SWE to snow disappearance date.

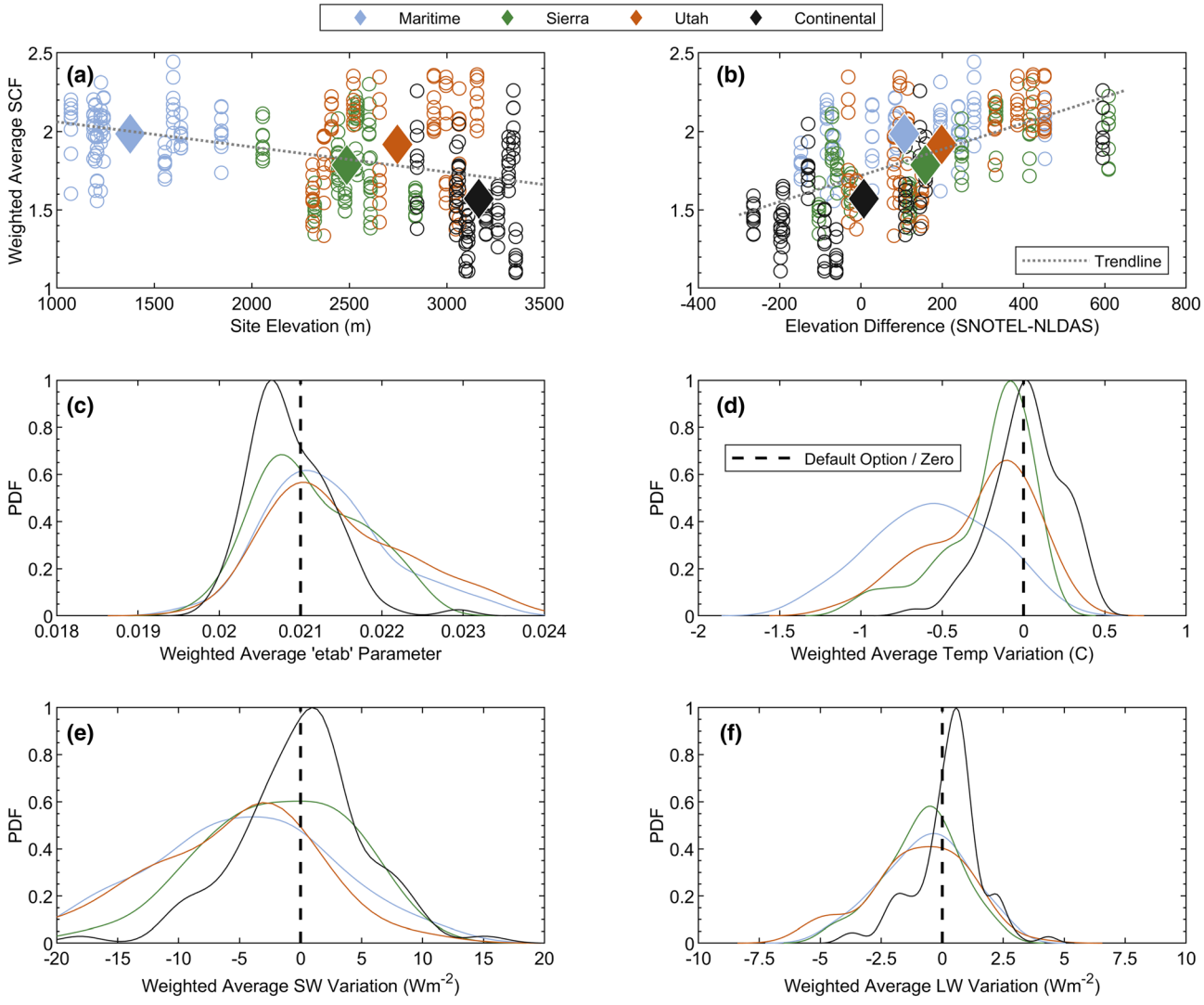

Figure 5. Comparison of weighted snowfall correction factor (SCF) by climate and site elevation (a) and by site elevation difference from corresponding NLDAS grid cells (b). In general, sites with lower elevation require a larger SCF adjustment from the PF to match observed depth. Although larger SNOTEL-NLDAS elevation differences require greater SCF corrections, all SCFs are above 1. Panels (c)–(f) show histograms of PF weighted average adjustments to other model inputs/parameters, showing that the PF does not consistently shift other perturbation sources to match observations. All additive perturbations are positive, that is, the perturbed value is the observed value + perturbation.

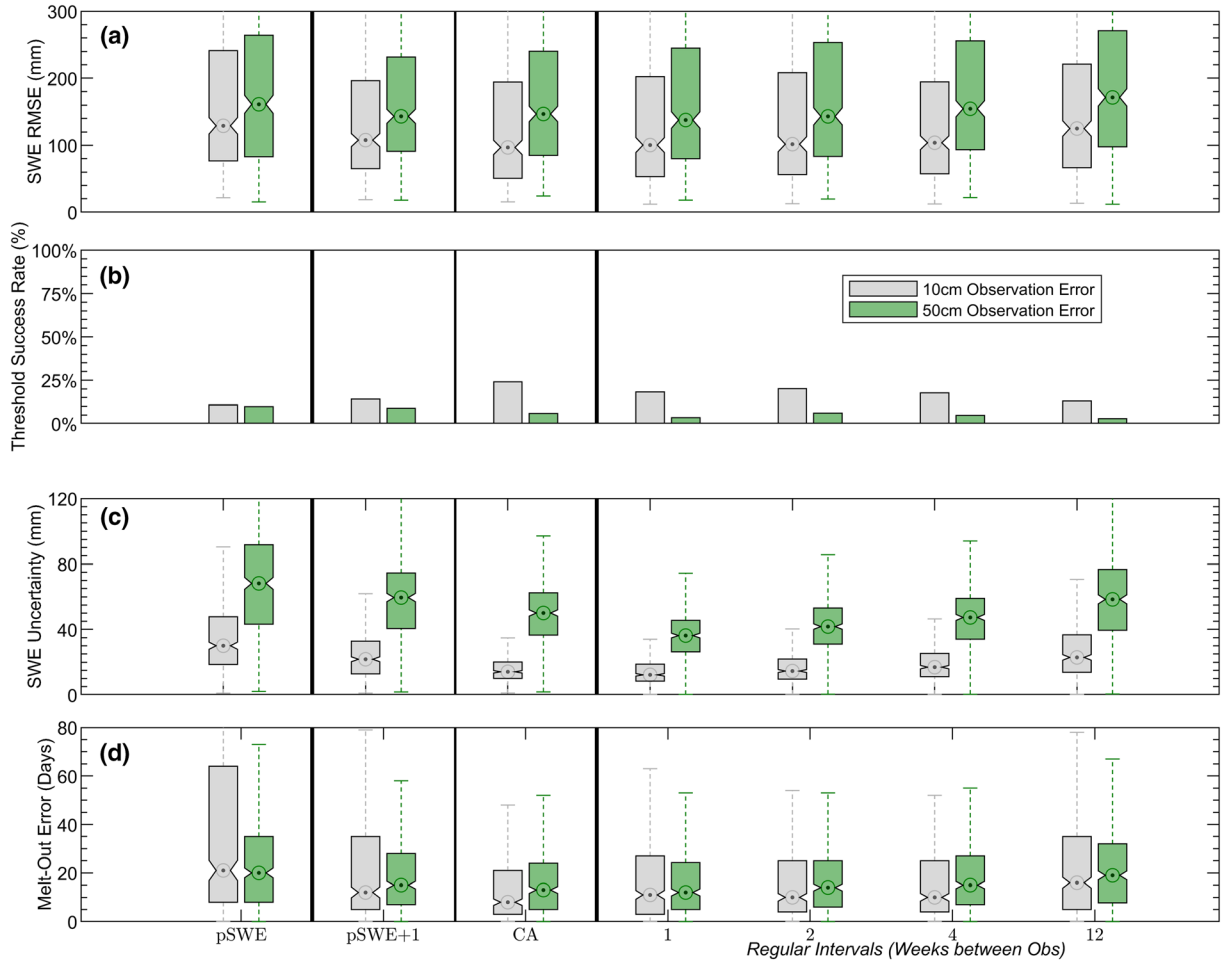

Figure 6. Box-and-whisker plots of (a) SWE RMSE, (c) SWE uncertainty, and (d) melt-out date error for all sites and years in different temporal sampling scenarios. For reference, the median open loop SWE RMSE in (a) is 254 mm and error in (d) is 40 days. Panel (b) shows the percent of simulations (among the 49 sites and 9 WYs) where the SWE RMSE meets the stated goal of 10% relative SWE error (i.e., relative to peak SWE at that given site/year).

Table 3
Comparison of SWE RMSE for Different Timing Scenarios, Precipitation Regimes, and Climates, Assuming 10 cm Observation Sampling Error

SWE RMSE (mm) 10 cm obs error	OL	pSWE	pSWE+1	CA	Regular intervals (weeks between obs) 1 week	2 weeks	4 weeks	12 weeks
All sites and years	254	129	108	97	100	102	104	125
Div. by peak SWE	*40%*	*20%*	*17%*	*15%*	*16%*	*16%*	*16%*	*20%*
Precipitation								
Wet years	355	187	173	152	155	162	167	186
Dry years	179	98	87	71	74	75	75	85
Climates								
Maritime—Pacific NW	483	291	217	221	230	235	235	247
Intermountain—Sierra	205	121	101	91	92	100	104	105
Intermountain—Utah	192	93	87	76	74	74	80	96
Continental—Colorado	148	87	65	51	52	56	64	83

Note. Wet and dry years are defined (for each site) as the years with the three highest and three lowest precipitation totals, respectively.

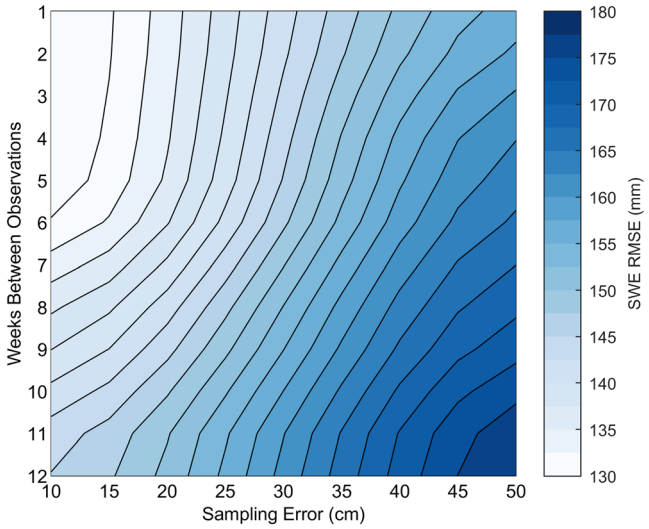

Figure 7. The full tradespace between observation timing (vertical axis) and sampling error (horizontal axis) endpoints discussed in the results. Contoured colors indicate SWE accuracy (RMSE). For ease of interpretation, only regular intervals are plotted here (i.e., regular weeks between observations).

our stated goal of 10% SWE error. Figure 6b shows the percent of sites/years where SWE RMSE meets this 10% error threshold (relative to peak SWE at the given site/year). For the 10 cm sampling error case, the "success rate" of meeting the stated RMSE goal ranges from 11% (pSWE observation timing) to 24% of sites/years (ASO timing). Meanwhile, the "success rates" for the 50 cm sampling error case is only 3–10% across sites/years.

Figure 7 shows SWE error for the full space of sampling errors between the two endpoints (10 and 50 cm) and for all regular timing intervals between 1 and 12 weeks. Although the axis units are not directly comparable, the contours in Figure 7 become more horizontal moving from left to right across the plot, indicating that SWE accuracy is more sensitive to observation timing for higher sampling errors (e.g., 50 cm compared to 10 cm). However, this is only true in terms of absolute RMSE differences, not relative (percentage) differences: The percentage change in SWE RMSE across timing scenarios for each sampling error case is consistent across the plot. Therefore, the conclusions we draw when analyzing the two endpoint cases (10 and 50 cm sampling error) are generalizable to other sampling scenarios.

SWE uncertainty clearly decreases with more frequent observations (Figure 6b). More frequent observations restrict the ensemble of particle estimates from spreading, thus lowering uncertainty. Furthermore, SWE uncertainty is greater with larger sampling error in snow depth (i.e., 50 vs. 10 cm). With a larger sensor error, particles that are farther from observations receive higher weights, increasing the weighted standard deviation measure of uncertainty. Still, these weights are applied nearly evenly above and below each observation, and the best particles receive the highest weights, so sensor error does not affect SWE accuracy (bias, RMSE) as much as it changes SWE uncertainty.

The accuracy of melt-out date estimation is improved with more observations (Figure 6c, Table S2). All timing scenarios are significantly better than the OL (median 40 days error), so again only one observation at peak SWE yields a majority of the possible improvement. Higher sensor error increases melt-out day errors, but still with improvements over the OL estimates (Table 2).

5. Discussion and Conclusions

The PF improves simulations of snow depth, density, SWE, and melt-out date across nearly all years and climates, supporting and extending the results of Smyth et al. (2019), which focused on a single site in California over 4 years. DA improves SWE estimates in two ways: (1) directly adjusting depth and (2) indirectly improving density, because depth and density are linked via densification processes such as overburden compaction. More accurate SWE yields improvements in melt-out date, an expected result given the well-established link between SWE accumulation and melt-out timing (e.g., Trujillo & Molotch, 2014). The absolute magnitude of SWE errors (both OL and PF) is higher in lower-elevation, maritime climates (e.g., Pacific Northwest) compared to higher-elevation, continental (e.g., Colorado) climates, but the relative improvement (from OL to PF) is similar in every climate (Table 3). Prior work has focused on the western slope of the Sierras (e.g., Tuolumne Basin), which is not included in our sample but should fall within the range of site conditions considered here; thus, our results can be compared to prior studies.

The PF weights simulations with precipitation inputs that exceed NLDAS by 50% or more. This large adjustment to precipitation inputs suggests that the magnitude of NLDAS precipitation in mountains is too low, consistent with previous studies (e.g., Enzminger et al., 2019; He et al., 2019; Henn et al., 2018). We use gridded NLDAS data in this study to more closely replicate the model inputs used in a distributed application. However, gauges at the SNOTEL sites record ~30% more precipitation than NLDAS (Figure S2), similar to the average difference across the western United States found by He et al. (2019). If we instead used precipitation measured at SNOTEL sites, the SCF values determined via DA would be approximately 1.2. An

adjustment of 20% is consistent with expected errors due to gauge undercatch (Rasmussen et al., 2012) and preferential snow drifting onto SNOTEL sites (Meyer et al., 2012).

An elevation-based precipitation adjustment (i.e., downscaling) would not be sufficient to match observations. While PF-weighted average SCF values are correlated with SNOTEL-NLDAS elevation differences, sites with no elevation difference still require SCFs of ~1.5X (Figure 5b). To confirm this result, we performed a separate downscaling test with precipitation inputs adjusted before assimilation-related perturbations were applied (Figure S1). Precipitation adjustments were based on the difference between the local (SNOTEL) and NLDAS grid cell elevations and a lapse rate of 0.02% increase in precipitation per meter of elevation increase, based on average precipitation lapse rates applied in central and western U.S. regions in PRISM (Daly et al., 2008). The precipitation adjustment eliminates the slope in the trendline in Figure 5b, that is, SCF differences between sites related to elevation differences. However, 75% of DA-weighted SCF values still ranged between 1.5X and 2.0X, indicating that downscaling-style adjustments alone are not sufficient to match observations. We repeated the downscaling experiment and similarly lapsed temperature (along with precipitation) in the test. SWE estimates were virtually unchanged, demonstrating that forcing perturbations during the assimilation process encompass potential downscaling changes.

While simply increasing precipitation forcing by the median SCF (~1.6X), for example, would improve the OL performance in many cases, PF performance would remain the same (see experiments with forcing data quality in Smyth et al., 2019). Further, the optimal bias correction factor is not known a priori. More importantly, the SCF varies considerably across sites/years (Figures 5a and 5b), and thus applying just a single value of SCF without DA would yield considerable errors.

Aside from precipitation, the PF does not consistently shift other model inputs or the most sensitive density-related parameter (Figure 7), suggesting that, of the factors we vary, precipitation is the largest source of model error—and model calibration would not be sufficient to reduce errors. Similar results were demonstrated for models that predict SWE directly based on meteorological variables and snow process representation (Günther et al., 2019; Raleigh et al., 2015), but to our knowledge, this result has not been previously shown for snow bulk density prediction. This result is supported by the SWE sensitivity test with FSM2 (Figure S4). Perturbations to particles in this study were based on existing literature (see section 3.5). Regardless, our results are not sensitive to particle perturbation ranges (Figure S5). One exception to our results is with temperature in the maritime zone, where the PF favors particles with lower temperature inputs. This is likely due to rain/snow partitioning in a mild winter climate, as DA attempts to increase snowfall to match observations.

The results indicate that the optimal timing/frequency and error in snow depth observations depend on the snow metric of interest. If the goal is to reduce errors in SWE, a DA framework can likely be successful with only a single snow depth observation. This result is consistent with Margulis et al. (2019). A single update can efficiently weight the particles with better precipitation inputs, compaction parameters, and other factors necessary to improve modeled depth, and thus density and SWE. In an operational sense, this result indicates that frequent, potentially expensive, observations of snow depth are largely unnecessary to reduce SWE errors, at least in the snow regime sampled by SNOTEL sites. This result is consistent with previous conclusions based on analysis of depth (e.g., Margulis et al., 2019), but here, we additionally validate density and SWE, across climates and in multiple years. We also move beyond the case of airborne lidar retrievals (e.g., Hedrick et al., 2018; Margulis et al., 2019) and demonstrate similar RMSE results with higher depth sampling error (50 vs. 10 cm), which are germane to current satellite photogrammetry capabilities. The accuracy of melt-out date estimates, which has implications for ecological studies and snowmelt timing applications, is more sensitive than SWE RMSE to sampling frequency of snow depth—but still, one observation yields most of the improvement over the OL.

However, if the goal is to reduce errors in SWE uncertainty, then more frequent and/or more precise snow depth observations are needed. This is an important consideration when evaluating the operational utility of the combined sensor, model, and DA system. More depth observations yield lower uncertainty, because they constrain the spread between particles during the WY and therefore lower our weighted standard deviation metric (Figure 1). In an operational framework, it is also important to recall that PF weights are carried backward in time through the interval preceding assimilated observations. Therefore, additional observations are

also essential to limit occurrences of long intervals between observations, when particles have not yet been reweighted.

If maximizing SWE accuracy and minimizing SWE uncertainty are both priorities, a platform with regular, approximately monthly (4-week) observations may be sufficient, even in climatologically abnormal years. This timing can be realistically achieved by satellite or targeted airborne platforms.

The ideal number/timing of observations in the mission design tradespace also depends on the sensor sampling error. For example, SWE RMSE is similar for weekly, 50 cm sampling error versus 12-week, 10 cm sampling error (Figure 6a). In this comparison, the 12-week case also has lower SWE uncertainty. This is an important consideration given the range in uncertainties of sensors that could be deployed to retrieve snow depth.

Here, we used the PF to assimilate depth observations at SNOTEL sites, yet the end goal is to utilize spatially extensive depth observations from remote sensing platforms. In our experimental setup (with validation from SNOTEL snow pillows and depth sensors), model performance would almost necessarily be worse if we assimilated "real" remotely sensed depth data, because the assimilated depth would not match SNOTEL measurements.

It is computationally feasible to apply the PF in applications which require simulation of large number of pixels, either when using high-spatial resolution lidar data from a single basin or lower-resolution global data. For reference, FSM2 can simulate 1 year at a point on one processor in approximately 2 s. To further increase computation speed, other DA techniques such as the particle batch smoother (PBS) can assimilate multiple observations at once and are more efficient (Margulis et al., 2015). We repeated the experiments here using the PBS and found similar results (Table S4).

None of the timing/sampling error scenarios we consider reach the 10% SWE error threshold articulated in the Decadal Survey at all sites and years (Figure 6a). This has some important implications for mission design, because our results imply we cannot achieve 10% SWE error at all sites/years by increasing the sampling frequency of snow depth alone. Given that SWE errors are more sensitive to sampling error than timing, one path forward is to target improvements in retrieval uncertainty (i.e., reduce snow depth error to values less than 10 cm). This may be difficult given budgetary and technological constraints, especially when moving from airborne/drone platforms to satellites. Instead, we suggest that improvements to modeled density are paramount to achieving the 10% targeted SWE error, as many large density estimation errors are evident, even with assimilation (Figure 3). The three-layer FSM2 model version implemented here tends to underestimate density. One path to produce more realistic snow density would be to increase the number of model snow layers, because previous research suggests more layers increases densification (Weiss, 2019).

In conclusion, this study illustrates the gains offered by additional or more precise snow depth observations, in the context of improving SWE estimates with DA. The optimal timing and sampling error of depth retrievals for a combined observation-model-DA framework depends on the maximum acceptable error and uncertainty. The snow and earth observation communities are currently articulating requirements for the accuracy and uncertainty of SWE retrievals (e.g., "global SWE ... every 3–5 days, to 10% accuracy for SWE values to 1 m" in the Decadal Survey). Our results highlight important considerations for achieving these goals. Future research should examine the effects of spatial resolution on depth assimilation for SWE estimation, along with the impact of sensor-specific limitations such as cloud and forest cover.

Acknowledgments
This work was supported by the National Aeronautics and Space Administration (NASA) Terrestrial Hydrology Program under Grant No. NNX17AL41G and NASA FINESST Grant No. 80NSSC19K1374. We thank Richard Essery for making FSM2 code available (https://github.com/RichardEssery/FSM2). NLDAS-2 forcing data were acquired through NASA GES DISC (https://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php), and SNOTEL data were acquired through the USDA NRCS (https://www.wcc.nrcs.usda.gov/snow/). This work utilized resources from the University of Colorado Boulder Research Computing Group, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. NLDAS model inputs, SNOTEL validation, and DA outputs from this study are available for free at CU Scholar: https://doi.org/10.25810/n1rc-zc38.

References

Anderson, E. A. (1976). A point energy and mass balance model of a snow cover. NOAA Technical Report NWS 19, 150.

Bales, R. C., Molotch, N. P., Painter, T. H., Dettinger, M. D., Rice, R., & Dozier, J. (2006). Mountain hydrology of the western United States. *Water Resources Research*, *42*, W08432. https://doi.org/10.1029/2005WR004387

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

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., et al. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, *29*(3), 317–319. https://doi.org/10.1002/joc

Deems, J. S., Painter, T. H., & Finnegan, D. C. (2013). Lidar measurement of snow depth: A review. *Journal of Glaciology*, *59*(215), 467–479. https://doi.org/10.3189/2013JoG12J154

Dozier, J. (2011). Mountain hydrology, snow color, and the fourth paradigm. *EOS, Transactions, American Geophysical Union*, *92*(43), 373–374. https://doi.org/10.1029/2011EO430001

Enzminger, T. L., Small, E. E., & Borsa, A. A. (2019). Subsurface water dominates Sierra Nevada seasonal hydrologic storage. *Geophysical Research Letters*, *46*, 11993–12001. https://doi.org/10.1029/2019GL084589

Essery, R. (2015). A factorial snowpack model (FSM 1.0). *Geoscientific Model Development*, *8*(12), 3867–3876. https://doi.org/10.5194/gmd-8-3867-2015

Feng, X., Sahoo, A., Arsenault, K., Houser, P., Luo, Y., & Troy, T. J. (2008). The impact of snow model complexity at three CLPX sites. *Journal of Hydrometeorology*, *9*(6), 1464–1481. https://doi.org/10.1175/2008JHM860.1

Garnaud, C., Bélair, S., Carrera, M. L., Derksen, C., Bilodeau, B., Abrahamowicz, M., et al. (2019). Quantifying snow mass mission concept trade-offs using an observing system simulation experiment. *Journal of Hydrometeorology*, *20*(1), 155–173. https://doi.org/10.1175/JHM-D-18-0241.1

Gleason, K. E., Nolin, A. W., & Roth, T. R. (2017). Developing a representative snow-monitoring network in a forested mountain watershed. *Hydrology and Earth System Sciences*, *21*(2), 1137–1147. https://doi.org/10.5194/hess-21-1137-2017

Grünewald, T., Schirmer, M., Mott, R., & Lehning, M. (2010). Spatial and temporal variability of snow depth and SWE in a small mountain catchment. *The Cryosphere Discussions*, *4*(1), 1–30. https://doi.org/10.5194/tcd-4-1-2010

Günther, D., Marke, T., Essery, R., & Strasser, U. (2019). Uncertainties in snowpack simulations—Assessing the impact of model structure, parameter choice, and forcing data error on point-scale energy balance snow model performance. *Water Resources Research*, *55*, 2779–2800. https://doi.org/10.1029/2018WR023403

He, C., Chen, F., Barlage, M., Newman, A., Tang, W., & Ikeda, K. (2019). Can convection-permitting modeling provide decent precipitation for offline high-resolution snowpack simulations over mountains? *Journal of Geophysical Research-Atmospheres*, *124*, 12631–12654. https://doi.org/10.1029/2019JD030823

Hedrick, A. R., Marks, D., Havens, S., Robertson, M., Johnson, M., Sandusky, M., et al. (2018). Direct insertion of NASA Airborne Snow Observatory-derived snow depth time-series into the iSnobal energy balance snow model. *Water Resources Research*, *54*, 8045–8063. https://doi.org/10.1029/2018WR023190

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Lundquist, J. D. (2018). An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology*, *556*, 1205–1219. https://doi.org/10.1016/j.jhydrol.2017.03.008

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, *5*(1), 1–25. https://doi.org/10.1080/10618600.1996.10474692

Kojima, K. (1967). Densification of seasonal snow cover. In *Physics of Ice and Snow, Proeedings of an International Conference on Low Temperature Science*, (pp. 929–952). Sapporo, Japan: Institute of Low Temperature Science, Hokkaido University.

Lievens, H., Demuzere, M., Marshall, H., Reichle, R. H., Brangers, I., de Rosnay, P., et al. (2019). Snow depth variability in the northern hemisphere mountains observed from space. *Nature Communications*, *10*(1), 4629. https://doi.org/10.1038/s41467-019-12566-y

Magnusson, J., Winstral, A., Stordal, A. S., Essery, R., & Jonas, T. (2017). Improving physically based snow simulations by assimilating snow depths using the particle filter. *Water Resources Research*, *53*, 1125–1143. https://doi.org/10.1002/2016WR019092

Margulis, S. A., Fang, Y., Li, D., Lettenmaier, D. P., & Andreadis, K. (2019). The utility of infrequent snow depth images for deriving continuous space-time estimates of seasonal snow water equivalent. *Geophysical Research Letters*, *46*, 5331–5340. https://doi.org/10.1029/2019GL082507

Margulis, S. A., Girotto, M., Cortés, G., & Durand, M. (2015). A particle batch smoother approach to snow water equivalent estimation. *Journal of Hydrometeorology*, *16*(4), 1752–1772. https://doi.org/10.1175/JHM-D-14-0177.1

Marti, R., Gascoin, S., Berthier, E., De Pinel, M., Houet, T., & Laffly, D. (2016). Mapping snow depth in open alpine terrain from stereo satellite imagery. *The Cryosphere*, *10*(4), 1361–1380. https://doi.org/10.5194/tc-10-1361-2016

Meyer, J. D. D., Jin, J., & Wang, S. Y. (2012). Systematic patterns of the inconsistency between snow water equivalent and accumulated precipitation as reported by the snowpack telemetry network. *Journal of Hydrometeorology*, *13*(6), 1970–1976. https://doi.org/10.1175/JHM-D-12-066.1

Mizukami, N., & Perica, S. (2008). Spatiotemporal characteristics of snowpack density in the mountainous regions of the Western United States. *Journal of Hydrometeorology*, *9*, 1416–1426. https://doi.org/10.1175/2008jhm981.1

Moller, D., Andreadis, K. M., Bormann, K. J., Hensley, S., & Painter, T. H. (2017). Mapping snow depth from Ka-band interferometry: Proof of concept and comparison with scanning Lidar retrievals. *IEEE Geoscience and Remote Sensing Letters*, *14*(6), 886–890. https://doi.org/10.1109/LGRS.2017.2686398

Molotch, N. P., & Bales, R. C. (2005). Scaling snow observations from the point to the grid element: Implications for observation network design. *Water Resources Research*, *41*, W11421. https://doi.org/10.1029/2005WR004229

National Academies of Sciences, Engineering, and Medicine (2018). *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space*. Washington, DC: The National Academies Press.

Nolin, A. W. (2010). Recent advances in remote sensing of seasonal snow. *Journal of Glaciology*, *56*(200), 1141–1150. https://doi.org/10.3189/002214311796406077

Painter, T. H., Berisford, D. F., Boardman, J. W., Bormann, K. J., Deems, J. S., Gehrke, F., et al. (2016). 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. https://doi.org/10.1016/j.rse.2016.06.018

Raleigh, M. S., Lundquist, J. D., & Clark, M. P. (2015). Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework. *Hydrology and Earth System Sciences*, *19*(7), 3153–3179. https://doi.org/10.5194/hess-19-3153-2015

Raleigh, M. S., Livneh, B., Lapo, K., & Lundquist, J. D. (2016). How does availability of meteorological forcing data impact physically-based snowpack simulations? *Journal of Hydrometeorology*, *17*(1), 99–120. https://doi.org/10.1175/JHM-D-14-0235.1

Raleigh, M. S., & Small, E. E. (2017). Snowpack density modeling is the primary source of uncertainty when mapping basin-wide SWE with lidar. *Geophysical Research Letters*, *44*, 3700–3709. https://doi.org/10.1002/2016GL071999

Rasmussen, R., Baker, B., Kochendorfer, J., Meyers, T., Landolt, S., Fischer, A. P., et al. (2012). How well are we measuring snow: The NOAA/FAA/NCAR winter precipitation test bed. *Bulletin of the American Meteorological Society*, *93*(6), 811–829. https://doi.org/10.1175/BAMS-D-11-00052.1

Serreze, M. C., Clark, M. P., Armstrong, R. L., McGinnis, D. A., & Pulwarty, R. S. (1999). Characteristics of the western United States snowpack from snowpack telemetry (SNOTEL) data. *Water Resources Research*, *35*(7), 2145–2160. https://doi.org/10.1029/1999WR900090

Smyth, E. J., Raleigh, M. S., & Small, E. E. (2019). Particle filter data assimilation of monthly snow depth observations improves estimation of snow density and SWE. *Water Resources Research*, *4*, 1296–1311. https://doi.org/10.1029/2018WR023400

Sturm, M. (2015). White water: Fifty years of snow research in WRR and the outlook for the future. *Water Resources Research*, *51*, 4948–4965. https://doi.org/10.1002/2015WR017242

Sturm, M., & Holmgren, J. (1998). Differences in compaction behavior of three climate classes of snow. *Annals of Glaciology*, *26*, 125–130. https://doi.org/10.3189/1998AoG26-1-125-130

Sturm, M., Taras, B., Liston, G. E., Derksen, C., Jonas, T., & Lea, J. (2010). Estimating snow water equivalent using snow depth data and climate classes. *Journal of Hydrometeorology*, *11*(6), 1380–1394. https://doi.org/10.1175/2010JHM1202.1

Trujillo, E., & Molotch, N. P. (2014). Snowpack regimes of the Western United States. *Water Resources Research*, *50*, 4149–4162. https://doi.org/10.1002/2013WR015173

van Leeuwen, P. J. (2009). Particle filtering in geophysical systems. *Monthly Weather Review*, *137*(12), 4089–4114. https://doi.org/10.1175/2009MWR2835.1

Vander Jagt, B., Lucieer, A., Wallace, L., Turner, D., & Durand, M. (2015). Snow depth retrieval with UAS using photogrammetric techniques. *Geosciences*, *5*(3), 264–285. https://doi.org/10.3390/geosciences5030264

Weiss, S. (2019). Snow grain evolution in forested and open regions—Implications for remote sensing. Oregon State University.

Wetlaufer, K., Hendrikx, J., & Marshall, L. (2016). Spatial Heterogeneity of Snow Density and Its Influence on Snow Water Equivalence Estimates in a Large Mountainous Basin. *Hydrology*, *3*(1), 3. https://doi.org/10.3390/hydrology3010003

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research*, *117*, D03109. https://doi.org/10.1029/2011JD016048