An Evaluation of Snow Initializations in NCEP Global and Regional Forecasting Models[Ⓞ]

NICHOLAS DAWSON, PATRICK BROXTON, XUBIN ZENG, AND MICHAEL LEUTHOLD

Department of Atmospheric Sciences, The University of Arizona, Tucson, Arizona

MICHAEL BARLAGE

Research Applications Laboratory, Boulder, Colorado

PAT HOLBROOK

Idaho Power Company, Boise, Idaho

(Manuscript received 13 November 2015, in final form 10 April 2016)

ABSTRACT

Snow plays a major role in land–atmosphere interactions, but strong spatial heterogeneity in snow depth (SD) and snow water equivalent (SWE) makes it challenging to evaluate gridded snow quantities using in situ measurements. First, a new method is developed to upscale point measurements into gridded datasets that is superior to other tested methods. It is then utilized to generate daily SD and SWE datasets for water years 2012–14 using measurements from two networks (COOP and SNOTEL) in the United States. These datasets are used to evaluate daily SD and SWE initializations in NCEP global forecasting models (GFS and CFSv2, both on $0.5^\circ \times 0.5^\circ$ grids) and regional models (NAM on $12\text{ km} \times 12\text{ km}$ grids and RAP on $13\text{ km} \times 13\text{ km}$ grids) across eight $2^\circ \times 2^\circ$ boxes. Initialized SD from three models (GFS, CFSv2, and NAM) that utilize Air Force Weather Agency (AFWA) SD data for initialization is 77% below the area-averaged values, on average. RAP initializations, which cycle snow instead of using the AFWA SD, underestimate SD to a lesser degree. Compared with SD errors, SWE errors from GFS, CFSv2, and NAM are larger because of the application of unrealistically low and globally constant snow densities. Furthermore, the widely used daily gridded SD data produced by the Canadian Meteorological Centre (CMC) are also found to underestimate SD (similar to GFS, CFSv2, and NAM), but are worse than RAP. These results suggest an urgent need to improve SD and SWE initializations in these operational models.

1. Introduction

Snow represents a primary source of water in many regions. For example, snowmelt from the Rocky Mountains accounts for the majority of the annual flow of the Colorado River in the southwestern United States (Palmer 1988; Serreze et al. 1999; Christensen and Lettenmaier

2007). The Colorado River provides up to an estimated \$1.4 trillion worth of economic activity for the entire basin, which was one-twelfth of the total U.S. gross domestic product in 2012 (based on calculations assuming an absence of river flow for the entire year of 2012; James et al. 2014).

Because of this economic dependence, forecasting the timing and magnitude of snowmelt is of critical importance. The accuracy of these forecasts relies critically on the estimates of snow on the ground at the time forecasts are made. For example, a forecast during a warm spell will predict far too little snowmelt if the amount of snow on the ground in the first place is too low. These “initial states” must be accurate in terms of snow cover and snow thickness.

A variety of observational data can be used to diagnose these initial states. Remotely sensed snow data are natural to use because they provide gridded estimates of snow

[Ⓞ] Supplemental information related to this paper is available at the Journals Online website: <http://dx.doi.org/10.1175/JHM-D-15-0227.s1>.

Corresponding author address: Nicholas Dawson, Department of Atmospheric Sciences, The University of Arizona, Physics–Atmospheric Sciences Bldg., Rm. 542, 1118 E. 4th St., Tucson, AZ 85721-0081.
E-mail: dawsonn@email.arizona.edu

quantities. In general, it is easier to detect the spectral characteristics related to the presence of snow than to the amount of snow (Frei et al. 2012). Therefore, the quality of remote sensing data is higher for snow products describing snow cover extent and lower for products describing mass of snow on the ground. Satellite estimates of snow depth (SD) and snow water equivalent (SWE), on the other hand, rely on passive microwave sensors, which suffer from signal quality issues from vegetation, reflection, and wet snow in mountainous terrain (Dietz et al. 2012; Mizukami and Perica 2012). Other retrieval methods such as light detection and ranging (lidar) and airborne gamma ray surveys provide high-quality estimates of snow variables (Grunewald et al. 2013; Kirchner et al. 2014), but temporal and spatial availability is limited (Pan et al. 2003), making their inclusion into large-scale (especially global) models difficult.

Station data [e.g., from the National Resources Conservation Service (NRCS) Snowpack Telemetry (SNOTEL) network and from the National Weather Services (NWS) Cooperative Observer Program (COOP) network in the United States] provide the most reliable information of SD and SWE. However, being point measurements, their inclusion with gridded data is not so straightforward. In some systems, they are merged with gridded data by determining biases between the point and gridded data, and interpolating those biases across the gridded fields (e.g., Brasnett 1999; Drusch et al. 2004).

Various snow interpolation techniques have also been developed to upscale point in situ data to produce gridded estimates of SWE and SD for model evaluations (e.g., Molotch and Bales 2005; Erxleben et al. 2002; Fassnacht et al. 2003; López-Moreno and Nogués-Bravo 2006; Dixon et al. 2014). For instance, Fassnacht et al. (2003) used satellite data to constrain their method of upscaling SNOTEL SWE data with regressions, but were unable to account for the nonlinearity of the data. While snow cover data are of relatively higher quality than SD or SWE estimates, daily visible satellite imagery is difficult to obtain because of the presence of clouds. Other upscaling methods (e.g., binary regression trees and multiple linear regressions) rely on the assumption that observations provide a representative sample for multiple predictors (e.g., Erxleben et al. 2002; Molotch and Bales 2005; Meromy et al. 2013).

This study aims to evaluate the initial states of SD and SWE in operational models at the National Centers for Environmental Prediction (NCEP) using upscaled in situ data. The NCEP models are evaluated because they provide operational forecasts that are used by a wide range of stakeholders who are interested in forecasts that involve snow. In addition, initializations of SD

and SWE in some of these models are largely based on uncertain SD data to begin with, and SWE is initialized from the SD data using a spatially and temporally constant snow density. Therefore, it is important to evaluate the effects of these uncertainties. In this study, we evaluate how well NCEP operational forecast models initialize SD and how severely the application of constant snow density adversely affects SWE initialization. A new snow density model (to replace the constant snow density assumption), evaluation of additional snow products, and a study of snow initialization impacts on seasonal forecasts will be presented in separate papers in the near future.

2. Data

a. Observations

This study uses daily data of SD and SWE from the SNOTEL network and SD from COOP. Together, the SNOTEL and COOP data represent a wide range of hydrometeorological conditions, and they provide the only consistently available daily observations of SD and SWE in the United States. Data are obtained for the 2012–14 water years (WYs), as this corresponds to the temporal overlap of the evaluated model initializations. Each WY begins on 1 October of the previous year and ends on 30 September of the current WY (e.g., WY 2012 spans from 1 October 2011 to 30 September 2012).

The SNOTEL network provides reliable snow data in remote mountainous locations. This study uses daily readings of SWE (measured using snow pillows) and SD (measured with ultrasonic SD sensors). Although the SNOTEL network provides reliable, fully automated, and quality-controlled data, there are some known deficiencies. For example, there can be measurement errors caused by temperature extremes (which influences the fluid within the snow pillow), ice bridging (where an ice layer suspends a snow mass above the pillow), and snow creep (the slow movement of snow downhill), which can all cause erroneous SWE readings (Serreze et al. 1999). SNOTEL data are also not representative of low-elevation areas, as SNOTEL sites are typically located high in the mountains. Because of this elevation bias, some studies have questioned the representativeness of SNOTEL data of the immediate surroundings (e.g., Molotch and Bales 2005; Meromy et al. 2013).

The NWS COOP network provides a high-density, long-term, daily dataset of SD that is representative of lower elevations, as most COOP stations are located near population centers. COOP data have been utilized by numerous studies (e.g., Brasnett 1999; Brown et al. 2001; Baxter et al. 2005; Ault et al. 2006; Slater and Clark 2006;

FIG. 1. Map of selected areas and locations of in situ data. Map background colors represent the elevation (m) and polygon colors indicate the type of observation.

Wi et al. 2012). As opposed to fully automated SNOTEL sites, trained volunteers manually collect data at each COOP station. As such, COOP data are more susceptible to temporal inconsistencies than SNOTEL data.

Each dataset is quality controlled to remove temporal inconsistencies in some of the time series. For SNOTEL sites, SWE time series are generally complete and continuous, though there are temporal inconsistencies in the SD time series. In general, many of these inconsistencies can be eliminated by ignoring large jumps in the data. This is also true for COOP SD data. In this study, for any SD (SWE) changes greater than 0.5 (0.2) m day^{-1} or reports of zero SD (SWE) when both the previous and subsequent days had nonzero values, the corresponding SD (SWE) was removed. We also removed sites where the reported elevation is obviously incorrect (i.e., if the reported elevation is outside of the possible range from -0.1 to 5 km based on elevation data for the United States). After these initial checks, data were also quality controlled visually to remove obvious errors (e.g., when large increments resulted in more than one day of erroneous SD) that were not removed by the above procedure (totaling 20 station days of data).

b. Study domains

This study evaluates the performance of forecasting models (described below) within eight $2^\circ \times 2^\circ$ boxes (whose locations are shown in Fig. 1). Boxes were selected based on elevation characteristics and availability of snow data. Also, each box is large enough to fully cover 16 complete grid cells from the coarsest-resolution models, which simplifies the averaging. Each box's elevation information was derived from the Shuttle Radar Topography Mission (SRTM; Farr et al. 2007) elevation data (Jet Propulsion Laboratory 2003). The elevation

data were resized from the native resolution of 3 arcs to approximately 0.01° (~ 1 km) resolution with a total of 40 000 pixels per box.

Selected boxes have a variety of elevation ranges, means, and standard deviations (Table 1). Boxes with standard deviations below 300 m are classified as “flat” and above this threshold as “mountainous.” Elevations span from -15 to 4327 m, with the gridbox means (standard deviations) ranging from 332 (56) to 2577 (521) m. The relatively flat Wisconsin (WI) and middle of Alaska [Alaska Mid (AKM)] boxes have low mean elevations, elevation ranges, and standard deviations. The next group of boxes [Colorado (CO), Montana (MT), and Yellowstone (YS)] is more mountainous, and all have similar standard deviations of approximately 380–390 m. The last group of boxes [Alaska South (AKS), Washington (WA), and Idaho (ID)] is the most mountainous and has standard deviations of about 510–520 m.

Each box has at least five observations per day during the study period except Alaska Mid, which has four SWE observations on 36 days. The average maximum number of reporting SNOTEL sites (for WYs 2012–14) per box (excluding Wisconsin, which has no SNOTEL sites) is 21 throughout winter–spring (from 1 December to 1 June; Table 1). The maximum number of reporting sites is used because of the intermittent dropout of observations. Colorado has the most (43) and Alaska Mid has the least (7). The average maximum number of daily reporting COOP sites per box is 17 throughout winter/spring. Wisconsin has the highest maximum number (39) and Alaska Mid has the lowest (3). COOP locations are, on average, 313 m lower than the average elevation in the selected boxes, which is consistent with the low-elevation bias of reporting stations identified in previous studies (e.g., Brasnett 1999).

TABLE 1. Elevation and observation metadata for selected boxes (shown in Fig. 1) sorted by standard deviation of 0.01° elevations. Observation counts are the maximum number of daily reporting stations from 1 Dec to 1 Jun across all three WYs. The elevation range is the difference between the highest and lowest elevation within each $2^\circ \times 2^\circ$ box.

	Elev range (m)	Elev mean (m)	Elev std dev (m)	COOP count	COOP elev	SNOTEL count	SNOTEL elev	Combined obs elev
WI	371	332	56	39	308	0	—	308
AKM	1389	525	266	3	257	7	763	510
CO	2605	2577	378	17	2374	43	2942	2658
MT	2216	1527	382	11	1035	14	1763	1399
YS	2698	2522	393	18	2037	26	2513	2274
AKS	2376	513	509	11	108	13	437	272
WA	4342	588	515	18	370	28	1241	804
ID	2413	1423	521	17	1016	18	1875	1445
Avg	2301	1251	378	17	938	21	1648	1209

c. Models

This study evaluates gridded snow initializations from global and regional operational forecast models that are used by NCEP. The models are the result of extensive research from the NCEP Environmental Modeling Center (EMC) and the NOAA Earth System Research Laboratory (ESRL). These models include the Global Forecast System (GFS; EMC 2003); the Climate Forecast System, version 2 (CFSv2, hereafter referred to as CFS; Saha et al. 2014); the North American Mesoscale Forecast System (NAM; Janjić and Gall 2012); and the hourly Rapid Update Cycle/Rapid Refresh model (RUC/RAP, hereafter referred to as RAP; Benjamin et al. 2004; McClung 2012, 2014). The Noah land surface model (LSM; version 2.7.1) is included in GFS, CFS, and NAM (with slight modifications for each model), while RAP includes the RUC LSM (versions were not tracked during this study period). LSM versions included in the Weather Research and Forecasting (WRF) Model, versions 3.4.1 and 3.5.1, were used for RAP and RAP, version 2, respectively (T. Smirnova 2016, personal communication). In addition, we evaluate the daily global SD product from the Canadian Meteorological Centre (CMC), which is widely used as ground truth in model evaluations (e.g., Niu and Yang 2007; Reichle et al. 2011; Yang et al. 2011; Liu et al. 2013; Kumar et al. 2014, 2015).

The NWS National Operational Hydrologic Remote Sensing Center (NOHRSC) Snow Data Assimilation System (SNODAS; Barrett 2003) is also included. SNODAS is self-described as the best guess for snow quantities in the continental United States (CONUS; Barrett 2003). The assimilation system uses downscaled RAP forcing data to drive a multilayered, uncoupled energy and mass balance model. SNODAS then assimilates all available snow measurements from ground, airborne, and satellite sources to nudge the model.

Some NCEP model initializations incorporate daily SD data from the Air Force Weather Agency (AFWA) Snow Depth Analysis Model (SNODEP) and the Interactive Multisensor Snow and Ice Mapping System (IMS; Helfrich et al. 2007). The IMS product consists of daily snow/ice cover maps, which are generated by analysts who interpret snow data from a variety of satellite and model products to generate operational snow cover maps (Helfrich et al. 2007). SNODEP, on the other hand, estimates snow thickness by combining satellite and surface observations (AFWA 2013). As a first guess, SNODEP uses the Special Sensor Microwave Imager/Sounder (SSMIS)-derived SD data, which are calculated based on a linear relationship with brightness temperature to SD (Foster and Davy 1988). The SD is constrained to 40 cm because of the unreliability of the algorithm for deeper snow packs (Northrop Grumman 2002). If the 2-m air temperature is above a tunable threshold, then SD is set to zero. Surface observations are then assimilated into the model based on distance and elevation thresholds and allow SD to exceed the 40-cm constraint of the remotely sensed data. Final processing includes manual adjustment by trained personnel, who compare SNODEP output to other satellite imagery before finalizing the data.

The NCEP models have various forms of initialization, including direct replacement of snow fields with the AFWA data, constraining SD fields with the AFWA data, and continuous cycling of snow variables from forecast to forecast. In addition, for the models that are initialized with the AFWA data, SWE is also initialized by multiplying SD by a constant snow density. For this study period, GFS SD is directly replaced with AFWA data (when available) and is masked with snow cover data from IMS. If IMS indicates snow cover, the greater of 5 cm or the AFWA SD is used as SD (EMC 2015). Subsequent initializations use the previous initialization's 6-h forecast for SD until a new AFWA

analysis is available. Initialized SWE is estimated by multiplying SD by a constant snow density of 100 kg m^{-3} (EMC 2015).

For NAM, snow initialization directly replaces modeled SD with AFWA data using the same rules applied to the GFS, but initialized SWE is, instead, estimated by multiplying SD by a constant snow density of 200 kg m^{-3} (EMC 2015).

For CFS, AFWA SD is not assimilated for the daily analysis if CFS SD is within one-half to twice the AFWA SD (Saha et al. 2010). If CFS SD is outside of this range, it is set to one-half or twice the AFWA data. Then, if no snow cover is present in the IMS product, snow is removed. If snow is present in IMS, a minimum SD of 2.5 cm is applied. As in GFS, a constant snow density of 100 kg m^{-3} is applied to obtain SWE (Saha et al. 2010), even if there is no adjustment to CFS SD. Note that even though these arbitrary constraints are different for some of the products (e.g., the GFS and NAM use a minimum SD of 5 cm if snow is present in IMS and the NAM uses a snow density of 200 kg m^{-3}), this evaluation includes all initialization constraints in order to evaluate products as they are used operationally.

Unlike the three operational models (GFS, CFS, and NAM), SD and SWE in RAP are cycled every hour (ESRL 2015). Cycled SD is masked by the IMS snow cover product twice daily (COMET 2015). Snow is removed if IMS indicates no snow cover and the model temperature is above freezing with no precipitation in the previous hour. Also, there is no adjustment of snow density over what is calculated by the model.

Assimilation of SD data is a little bit different in CMC. The CMC data use a temperature index snowmelt model for a first guess and then utilize an optimum interpolation technique to adjust the model forecast toward observed values (Brasnett 1999). Observations include land surface synoptic (SYNOP) reports available on the Global Telecommunication System, meteorological aviation reports (METARs, when available), and special aviation (SA) reports from the World Meteorological Organization (WMO) information system (Brown and Brasnett 2010). For this process, the background field is estimated based on the previous (6 h old) SD plus an estimated increment based on new precipitation, temperature, and snow density (which evolves through time). The background data are then bilinearly interpolated to observation points to find a bias at the observation points. Second-order autoregressive functions are then used to obtain the horizontal and vertical correlation matrices of errors between all pairs of observation points, as well as between the first guess and the observations. These matrices are then combined to determine an optimal weight matrix (Brasnett 1999), which is used to find an

interpolated bias field that is added to the background field to get the new analyzed field.

Model data are obtained from NOAA's National Operational Model Archive and Distribution System (NOMADS). The downloaded model data have various resolutions, which may be different than the native resolutions for each model. Downloaded GFS and CFS data are on a $0.5^\circ \times 0.5^\circ$ latitude–longitude grid. These data are simply subset for each of the $2^\circ \times 2^\circ$ boxes as 16 complete grid cells fit into each box. NAM model data are on a 12-km grid with a Lambert conformal projection and include a maximum (minimum) of 270 (265) grid cells per box. RAP data are on a 13-km grid with a Lambert conformal projection and include a maximum (minimum) of 222 (214) grid cells per box. The downloaded CMC data are in a polar stereographic projection with a grid resolution of 23.813 km at 60°N and include a maximum (minimum) of 87 (35) grid cells per box. SNODAS is available on a 30-arc-s ($\sim 1 \text{ km}$) latitude–longitude grid and each box includes 57 600 grid cells. For each of the $2^\circ \times 2^\circ$ boxes, NAM, RAP, CMC, and SNODAS data are subset by including model grid cells whose center latitude and longitude are within each box, as the number of grid cells per box is large.

CFS, GFS, NAM, and RAP are used at NCEP for operational regional and global forecasting. NAM and RAP have a higher resolution than GFS and CFS [whose native grid sizes are T574 ($\sim 27 \text{ km}$) and 0.25° , respectively, for this study period]. While the AFWA data are not freely available for download, direct insertion of the data into the GFS and NAM allows for comparison of AFWA at two different model grid resolutions. In addition, since RAP does not utilize AFWA data for its initialization, its inclusion allows comparison of a non-AFWA initialized model to models that utilize AFWA data. It needs to be emphasized that the snow initializations of these operational models, as described in this section, are valid for the data period (WYs 2012–14), as they evolve with time.

3. A new method to upscale point measurements to area average

a. Method to compute area-averaged snow depth

Prior to comparing the COOP and SNOTEL data with the model initialization data in the eight boxes, we upscale daily point observations to get area averages within the boxes (shown in Fig. 1) using our new upscaling method described here (a flowchart of the technique is provided in Fig. S1 in the supplemental material). Our method is based, in concept, on that of Fassnacht et al. (2003), which involves interpolating residuals from a relationship between SWE and

FIG. 2. SD regressions for the WA area (Fig. 1) using a linear fit (black) and PWBIN (red) for (a) 15 Jan and (b) 15 Apr 2012. (c) A time series of the area-averaged SD for each first-guess method.

elevation. The key difference is that we use a piecewise linear regression with elevation instead of the linear regression with elevation used in Fassnacht et al. (2003), and we use a different method of interpolation of the residuals.

For the first step in our method, all quality-controlled observations from both SNOTEL and COOP stations are combined into a single dataset for each $2^\circ \times 2^\circ$ box and then separated into 100-m elevation bins. Each binned SD value is determined as the median of all observations within the bin. The second step is to find a critical bin, which is the first bin with either zero SD or 5% of the maximum binned SD starting from the highest-elevation bin moving downward. Once the critical bin (usually with zero or small SD or SWE) is established, a least squares regression is applied for all bins at and above the critical bin. This regression is forced through the critical bin to avoid discontinuities with the next-lowest-elevation bin. A piecewise linear regression is applied to all bins at and below the critical bin. If the critical bin is the lowest-elevation bin, a single linear regression is applied to all bins. The third step is to extrapolate the value in the lowest (or highest) elevation bin to all elevations below (or above) this bin. These regressions [collectively called piecewise bins (PWBIN)] are applied to each 0.01° pixel to determine a first-guess field for each day.

Note that lidar data have shown a positive relationship of SD and elevation (i.e., increasing SD with increasing elevation) up to a certain elevation (~ 3300 m; Kirchner et al. 2014) followed by a negative relationship (i.e., decreasing SD with increasing elevation). The increase of observed SD and SWE with elevation in Fig. 2 is consistent with this finding as the highest elevation of sites in this box is less than 2000 m. It is unknown if such a critical elevation

exists everywhere and, if so, what the critical elevation is. Therefore, a constant extrapolation (above the highest-elevation bin) is chosen in our PWBIN method.

The fourth step is to interpolate the residuals between each observation and the first-guess field. Residuals are computed by taking the difference between the station data and the nearest 0.01° first-guess pixel. Numerous methods were developed in the past to upscale and/or interpolate residuals of point observations of various complexity, but validation of these methods becomes increasingly difficult with available surface observations for large-scale models. Here, three widely used interpolation methods are tested. The simplest method is to interpolate using inverse distance squared weighting (IDW), which gives more weight to observations closest to the analyzed pixel (Balk and Elder 2000; Erxleben et al. 2002; Fassnacht et al. 2003). The second method is based on kriging (KRIG; Erxleben et al. 2002) and involves 1) fitting a model to the data's semi-variogram (which describes the spatial variability of the data) and 2) performing the kriging. In this study, a spherical model is used to approximate the semi-variogram. The final method uses the optimum interpolation (OI) method outlined in Brasnett (1999). At its heart, this method relies on a second-order autocorrelation equation to determine the spatial correlation coefficient between two points:

$$\alpha(d_{ij}) = (1 + 0.018 \times d_{ij}) \times e^{(-0.018 \times d_{ij})} \times e^{[-(h_{ij}/800)^2]}. \quad (1)$$

Here, d_{ij} is the horizontal separation of points i and j (km) and h_{ij} is the vertical separation of the points (m). The constant coefficients are 0.018 km^{-1} and 800 m. This OI method is also utilized by two operational centers

(CMC and the European Centre for Medium-Range Weather Forecasts) for daily estimates of SD. Equation (1) is a function of both horizontal and vertical distances, and as such, this method is the only one (among the above three methods) that limits the influence of high (or low)-elevation observations on low (or high)-elevation areas, which is important in mountainous terrain.

Last, the interpolated biases are added to the first-guess field to determine the final SD field. Any negative SDs that result from adding a negative bias to a shallow snowpack are adjusted to zero for all methods. The above methods are applied to both SD and SWE fields, and their daily averages over the eight boxes in Fig. 1 are evaluated next to select the best method for upscaling point measurements to area averages.

b. Analysis of upscaled snow depth

Our new PWBIN regression method is able to more accurately fit the observations than a single linear regression. As an example, Figs. 2a and 2b show the single linear and PWBIN regressions for the SD first-guess field for the WA box for 15 January and 15 April 2012. Note that the single linear regression overestimates SD for elevations between approximately 300 and 1200 m on both days, which correspond to the accumulation (Fig. 2a) and melt (Fig. 2b) periods. The PWBIN method is able to capture the nonlinear relationship between high- and low-elevation observations in both cases.

To translate these errors at individual locations to area-averaged errors, we divide the box into $0.01^\circ \times 0.01^\circ$ pixels and apply the regressions in Figs. 2a and 2b to each pixel. Then the pixel-averaged values are used to represent the average for the box. Figure 2c shows area-averaged SD time series for the same (WA) box for the entire 2012 water year. For most of the winter, using a single linear regression causes first-guess estimates of area-averaged SD to be 10%–30% higher than using the PWBIN method. Large differences between the methods also exist from March to May. However, the two methods are similar in early winter.

Next, the interpolation methods (discussed in section 3a) of the residuals from the first-guess field are compared. Final results (after interpolated residuals are added to the PWBIN first-guess field) are shown in Figs. 3 and 4 for two days (15 January and 15 April 2012) for the WA box. Compared to the other two methods, the IDW method has more snow at lower elevations (where in situ measurements indicate less snow; Figs. 3a, 4a). This deficiency can be attributed to the fact that the IDW method does not consider vertical elevation changes and hence interpolates SD biases (at higher elevations) down to valley floors.

An attempt to address this deficiency was considered by Fassnacht et al. (2003), whereby residuals were regressed to 5000 m using a lapse rate of 9.8 mm km^{-1} . The lapsed residuals were then interpolated with IDW and lapsed back to the original elevations. However, we found that their elevation-detrended method using a single linear fit for the first guess did not significantly change the area average compared to a regular IDW approach (not shown). The addition of a snow cover mask via visible satellite imagery by Fassnacht et al. (2003) is unreliable for daily analysis here because of cloudy conditions, and it introduces additional uncertainty.

The KRIG method generally results in nonzero SD in fewer areas than the IDW method (i.e., more confined to higher elevations), but in more areas than the OI method (Figs. 3b, 4b). Like the IDW method, the KRIG method also ignores the vertical elevation differences, but it provides a more reasonable estimate than IDW because it accounts for correlation distances.

For the OI method, nonzero SD is mainly restricted to high elevations (Figs. 3c, 4c), which is probably most realistic. On average across all boxes and water years, the difference of area-averaged SDs based on the KRIG and OI methods is negligible compared to the actual SD.

To provide a more quantitative measure of the performance of each interpolation method, we performed a comparison between the final products (PWBIN with IDW, KRIG, and OI) and IMS snow cover data subset for the WA box. For this comparison, the IMS data were reprojected from a polar stereographic projection with 4-km grid spacing (true latitude of 60°N) to 0.01° pixels using nearest-neighbor interpolation for direct comparison with the upscaled SD observations. For each of the interpolation methods, all pixels with upscaled SD over 5 cm were classified as snow covered, and all other pixels were classified as snow free. The PWBIN–OI method has the highest agreement with the IMS data (Table S1 in the supplemental material; from 1 November 2011 to 1 May 2012). Percentages of agreement (disagreement) are calculated as the number of pixels in each category divided by the total number of pixels within the selected time period (7.32 million pixels). The PWBIN–OI method has the lowest percentage of false positives, a similar percentage of false negatives, and the highest combined agreement (77.3%) of the three residual interpolation methods. Threshold values of 0 (10) cm produce lower (higher) agreement, but the PWBIN–OI method still has the highest agreement regardless of the threshold value selected (Table S1 in the supplemental material).

Robustness of the PWBIN–OI method was then tested with a data denial test. For instance, seven SNOTEL and three COOP locations (approximately

FIG. 3. Final analyzed SD (cm) after addition of interpolated residuals to the PWBIN first-guess field for the WA area (Fig. 1) on 15 Jan 2012 (a) using the IDW method for interpolating the residuals with an area-averaged value of 26.3 cm, (b) utilizing the KRIG method with a mean of 30.0 cm, (c) using the OI method with a mean of 25.6 cm, and (d) SNOTEL SDs with a mean of 25.3 cm. Circles represent COOP and SNOTEL SDs.

20% of each dataset) were randomly removed from the box with the highest number of observations (the CO box) as a validation dataset for WY 2012. This removal of 10 locations produces a relatively small mean absolute error (MAE; over these 10 sites), with the temporal average of the ratio between daily SD MAE averaged over the 10 sites and the observed mean SD from 1 January to 1 April 2012 being 18.7% (Fig. S2 in the supplemental material). This time period was chosen as an estimate of initial snowpack to the date of maximum SWE. Furthermore, the temporal correlation of the estimated and observed mean SDs over the 10 sites is as high as 0.95.

Based on these evaluations, we use the PWBIN–OI method to compute the daily area-averaged SD for each box in Fig. 1 for model evaluations in section 4. The use of spatial averaging for model SD and SWE validation has been performed in previous studies (e.g., Pan et al. 2003; Kumar et al. 2014, 2015). Area averages are utilized here because of varying model output grid sizes (ranging from ~ 50 to 1 km) and the use of ~ 1 -km-resolution validation data. While a relatively coarse model grid size is not expected to capture small-scale variability of the snowpack, initialized SD and SWE should have enough skill to represent average conditions within each $2^\circ \times 2^\circ$ box.

FIG. 4. As in Fig. 3, but for 15 Apr 2012. The area-averaged SD is (a) 88.3, (b) 80.0, (c) 74.8, and (d) 80.5 cm.

c. Method to compute area-averaged snow water equivalent

The same piecewise regression methodology (i.e., PWBIN) can be applied to estimate SWE, which is available from the SNOTEL network only. Like the above process to estimate SD, estimating SWE involves 1) separating the point observations into 100-m elevation bins, 2) identifying a critical bin (above which a linear regression is applied and below which a piecewise linear regression is applied), and 3) interpolating the residuals between the observations and the first-guess field (using the OI method). In other words, we also use the PWBIN–OI method to compute the daily area-averaged SWE for each box in Fig. 1 for model evaluations in section 4.

4. Evaluation of snow initializations in NCEP operational forecast models

a. Validation of model snow depth initialization

The model initializations that we tested (except RAP) consistently underestimate the area-averaged SD by substantial amounts (Fig. 5), particularly in boxes over the CONUS (the only exception is the AKM box). In addition, initializations that utilize the AFWA SDs (GFS, CFS, and NAM) exhibit sporadic jumps in area-averaged SD. Most of these erroneous spikes occur before January 2012 (e.g., for the first few months of the Yellowstone and Colorado time series in Fig. 5), except for a spike in February 2014 for areas in the Rocky Mountains at the same time as the dramatic decrease in

FIG. 5. Area-averaged SD (cm) for each box (Fig. 1) for the PWBIN-OI method (black), CMC (blue), GFS (orange), CFS (turquoise), NAM (purple), RAP (dark green), and SNODAS (light green) for WYs 2012–14. The x axis indicates month and year. Note that the y axis range differs.

SD for RAP. Note that a sharp decline in area-averaged RAP SD is present the day after RAP was upgraded to a new version in 2014 for the MT, YS, ID, and CO boxes (see section 5 for further discussion). RAP also sustains small area-averaged SD values through the summer in the WA box, due to the very large SD value (e.g., up to 4283 mm on 1 August 2012) for the 13 km $\times$ 13 km grid cells covering Mt. Rainier, as well as other grid cells covering Mt. Adams (which are both permanently snow covered mountains). The large SD values in these and surrounding grid cells on this day are shown in Fig. S3 in the supplemental material.

Figure 5 shows that our PWBIN-OI method agrees very well with SNODAS, which assimilates all available snow measurements from ground, airborne, and satellite sources. Even though this is not an independent validation of our PWBIN-OI method because of the use of SNOTEL and COOP data in SNODAS, this agreement provides further confidence that our method produces reasonable results in the Alaskan boxes (as SNODAS is only available for the CONUS).

Area-averaged SDs from GFS, CFS, and NAM across all water years (from 1 December to 1 June) are similar for each box, and they are generally much lower than the upscaled observations (Table 2). RAP results show much smaller biases than these three models. GFS, CFS, and NAM also show similar MAEs that are also much

greater than those for RAP (Table 2). The similar performance of the regional model (NAM) and two global models (GFS and CFS) also indicates that gridcell size does not significantly impact results.

Performance of each NCEP model at the relatively “flat” AKM and WI boxes is higher than at the more mountainous boxes for all water years (Table 2). The two relatively flat areas have model initialized average SD MAEs of 3.6 and 12.3 cm (35% and 29% of the mean, respectively) as opposed to MAEs of 37.7–72.7 cm for all other boxes (ranging from 56% to 84% of the mean). The WA box has the highest average SD MAE (36.1 cm) of the initializations relative to the observed mean SD (43.2 cm). The poor performance in the mountainous boxes exemplifies the difficulties of these models in estimating snow quantities in these environments.

CMC’s performance is consistently worse than that of RAP (Table 2). CMC’s mean bias is slightly higher than the GFS, CFS, and NAM, while its MAE is comparable to those of these three models (Table 2). SNODAS MAE is approximately half of all other products for both SD and SWE, except for Wisconsin because of larger differences during March of WY 2013.

Additional statistics for maximum SD, date of maximum SD, and last day of SD over 5 cm (end of snow season) averaged for all WYs are provided in Table 3.

TABLE 2. The average daily SD (SWE) for all boxes (cm) from Fig. 1 is shown. The time period for the averages was from 1 Dec to 1 Jun for WYs 2012–14. Averages are also provided for the six boxes in the CONUS (i.e., excluding AKM and AKS) for direct comparison across all products. Also, the average daily SD (SWE) MAE between each model initialization and our area-averaged values (i.e., PWBIN–OI) for all boxes (cm) is shown. Note that no SWE data exist for WI because of the absence of SNOTEL data.

	GFS	CFS	NAM	RAP	CMC	SNODAS	PWBIN–OI
Mean SD (SWE)							
WI	8.5 (0.9)	7.9 (0.8)	7.8 (1.6)	7.4	7.7	11.8 (2.3)	10.3
AKM	32 (3.2)	37.0 (3.7)	—	—	26.2	—	42.4 (7.2)
CO	16.0 (1.6)	15.4 (1.5)	15.2 (3.2)	19.4	7.8	42.6 (10.8)	51.3 (16.6)
MT	13.3 (1.3)	13.1 (1.3)	13.2 (2.7)	54.7	12.7	79.1 (24.3)	73.1 (24.3)
YS	27.6 (2.8)	31.5 (3.1)	26.5 (5.5)	80.1	34.4	106.4 (29.1)	109.7 (35.6)
AKS	24.9 (2.5)	32.5 (2.9)	—	—	53.1	—	81.7 (25.6)
WA	4.8 (0.5)	5.8 (0.6)	4.9 (1.1)	20.1	1.5	44.3 (14.9)	43.2 (17.4)
ID	7.5 (0.7)	7.8 (0.8)	6.2 (1.3)	22.8	6.5	42 (12.1)	50.7 (16.4)
Avg without AKM and AKS	13.0 (1.3)	13.6 (1.3)	12.3 (2.6)	34.1	11.8	54.4 (15.6)	56.4 (22.1)
SD (SWE) MAE							
WI	4.3	3.4	4.4	2.8	2.9	9.1	—
AKM	13.7 (4.1)	7 (3.6)	—	—	16.2	—	—
CO	37.6 (14.9)	36.8 (15.0)	38.3 (13.7)	32.5	43.5	18.7 (7.8)	—
MT	59.9 (22.8)	60 (23.0)	60.2 (21.6)	27.7	60.4	20.1 (8.0)	—
YS	84.2 (32.7)	80.8 (32.5)	85.7 (30.4)	36.9	75.3	29.2 (12.5)	—
AKS	56.8 (23.1)	50.3 (22.8)	—	—	29	—	—
WA	38.5 (16.9)	37.4 (16.8)	38.5 (16.3)	24.6	41.7	15.2 (6.9)	—
ID	43.1 (15.6)	42.9 (15.6)	44.6 (15.1)	28.8	44.2	19.5 (7.2)	—
Avg without AKM and AKS	44.6 (20.6)	43.6 (20.6)	45.3 (19.4)	25.6	44.7	18.6 (8.5)	—

The average maximum SD and date of the maximum SD are susceptible to the erroneous early season spikes in initialized SD apparent in GFS, CFS, and NAM (Fig. 5). SNODAS and our PWBIN–OI results agree with each other very well (e.g., the average difference of the maximum SD is -1.3 cm only in Table 3). Compared with our PWBIN–OI results, RAP is the best performer for all of the additional statistics among the five products (GFS, CFS, NAM, RAP, and CMC). Snow season end date (excluding WA for RAP) is about 2 weeks early on average for RAP and over 1 month early for other products. Maximum SD is underestimated by all products, but again, RAP provides the lowest bias of -25.7 cm.

b. Validation of model snow water equivalent initialization

Similar to SD, initialized SWE is consistently underestimated by these model initializations (Fig. 6). Compared with the observed SWE, both the mean bias (in magnitude) and MAE are, on average, largest for GFS and CFS (Table 2, with SWE values shown in parentheses). Similar to SD, the two relatively flat boxes have smaller biases and MAEs compared to all other areas. The agreement between SNODAS and our PWBIN–OI results is much better than those between the other products and our results.

Errors in initialized SWE are directly related to SD errors because of the conversion of SD to SWE for GFS,

CFS, and NAM. However, the use of constant density (which is generally too low for snow that accumulates on the ground) exacerbates this underestimation. This can be seen by comparing the ratios between modeled and observed SD and SWE. For GFS, CFS, and NAM, this ratio is always higher for SD than for SWE. This can be seen in Table 4, which shows these ratios for the months of December (Table 4) and April (Table 4) averaged for all water years. December and April are chosen to characterize an accumulation period and a melt period, respectively.

For December, the average ratio between modeled and observed SDs for all boxes was similar for GFS, CFS, and NAM (Table 4), and the overall average ratio of SD across all of the boxes is 0.61. The similarity is not surprising because of the use of similar initialization data for the three models. Small variations between the models for individual boxes are likely to be the result of slight differences in initialization (for CFS) and grid size (for NAM). In contrast, the average ratio between modeled and observed SWEs from NAM is approximately twice as large as that from both GFS and CFS, largely because NAM SWE is initialized using a constant density of 200 kg m^{-3} , which is twice that of GFS or CFS. Ratios for SNODAS are closest to 1 for all three quantities.

A large decrease of both SD and SWE ratios occurs in April for all three NCEP initializations (Table 4), indicating the greater deficiency of these products during

TABLE 3. The maximum SD (cm), the Julian day of the WY for the maximum SD, and the snow season end date (Julian day of the WY when area-averaged SD < 5 cm) averaged across all WYs are shown. Note that the WA average on the last day of the snow season for RAP was removed because of persistent nonzero SD, as shown in Fig. 5. The averages over the six boxes in the CONUS (i.e., excluding AKM and AKS) are also provided.

	GFS	CFS	NAM	RAP	CMC	SNODAS	PWBIN-OI
Max SD							
WI	31.8	24	32.1	27.6	25.6	37.8	31.8
AKM	54.3	65.4	—	—	45.5	—	74.2
CO	57.7	39.4	55.2	49.3	25.2	79.7	91
MT	72	44	66.8	96.3	32.9	122.4	112.2
YS	118.6	83.2	106.4	135.8	77.9	161.7	162.4
AKS	52	64.9	—	—	91.2	—	124.5
WA	39	39.3	39.3	41.3	14.2	77.9	78
ID	34	29.3	28.5	53.3	24.5	70.7	80.3
Avg without AKM and AKS	58.9	43.2	54.7	67.3	33.4	91.7	92.6
Date of max SD							
WI	148	144	148	143	134	155	142
AKM	108	145	—	—	155	—	149
CO	92	129	101	147	147	162	170
MT	110	116	106	161	124	179	188
YS	135	121	135	159	157	174	176
AKS	133	149	—	—	149	—	167
WA	109	113	109	135	110	156	156
ID	128	116	121	147	115	147	169
Avg without AKM and AKS	120	123	120	149	131	162	167
Last Julian day of snow season							
WI	183	185	179	172	182	186	181
AKM	226	223	—	—	216	—	226
CO	207	204	206	215	202	245	244
MT	213	199	199	270	206	277	279
YS	228	231	222	275	234	268	268
AKS	232	228	—	—	288	—	252
WA	180	166	161	—	144	281	277
ID	188	181	181	228	168	255	257
Avg without AKM and AKS	200	194	191	232	189	252	251

the snowmelt period. GFS and CFS have similar average SD ratios in April, but the NAM SD ratio is twice as large. Without the WI, AKM, or AKS averages, SD ratios are 0.08, 0.06, 0.06, and 0.91 for GFS, CFS, NAM, and SNODAS, respectively. Again, NAM has a larger SWE ratio than the GFS and CFS in April because of its use of a constant density that is twice as large as GFS and CFS.

The effective snow densities can be computed as the area-averaged SWE divided by area-averaged SD. Then an effective average density ρ ratio between modeled and observed values can be calculated (Table 4). The effective average density ratio from NAM is roughly twice as large as that from both GFS and CFS, again because NAM SWE is initialized using a constant density of 200 kg m^{-3} , which is twice that of GFS or CFS. As the snow season progresses, the actual snow density increases because of compaction and snowmelt, and the constant density approximation worsens the initialization of SWE in these products. Therefore, there is a noticeable drop of ρ from December to April (Table 4).

c. Uncertainty quantification

An obvious question in the above model validations is how the area-averaged in situ data uncertainty based on the PWBIN-OI method would affect the conclusions. Here we address this question through in situ data denial tests.

For the CO box, 50% of sites are randomly denied for each day and the PWBIN-OI method is used to compute the daily SD. Then this process is repeated 100 times for each day. The daily ensemble mean SD (from the 100 iterations) is compared to the result without data denial (i.e., using all sites) to obtain MAEs for the period from 1 October 2011 to 1 June 2012 (as the model initialization average MAE is lowest in WY 2012 for the CO box). These MAEs (due to data denial) are then compared with model initialization MAEs discussed in sections 4a and 4b.

To make even more stringent tests, the above test is repeated for data denials of 50%, 75%, and 90% with denial of only COOP, only SNOTEL, and of the

FIG. 6. As in Fig. 5, but for area-averaged SWE (cm) for five datasets only.

combined (COOP–SNOTEL) data (Fig. S4 in the supplemental material). The highest MAE (7.8 cm; Table S2 in the supplemental material) occurs when 90% of only SNOTEL data are denied, because SNOTEL data outnumber COOP data in the CO box (Table 1). The lowest MAE (0.6 cm) occurs when 50% of the combined data are denied. Overall, MAEs from these data denial tests are substantially less than the model initializations MAE (26.7 cm), and hence the conclusions in sections 4a and 4b are insensitive to the number of in situ sites used.

Furthermore, of the three combined methods tested (PWBIN regression for the first-guess fields and three residual interpolation methods), the chosen method (PWBIN–OI) produced the lowest area-averaged SD for a majority of the days across all boxes (figure not shown). Therefore, no matter which combined method is used, the conclusions in sections 4a and 4b (i.e., NCEP model initializations substantially underestimate SD and SWE) are not affected.

5. Conclusions

A new method for upscaling SD and SWE from point SNOTEL and COOP measurements has been developed and used to validate operational NCEP model snow initializations (GFS, CFS, NAM, and RAP). The new PWBIN method is found to be superior to single linear regression estimates for first-guess fields of SD. Among

the three different methods to interpolate deviations from the first guess, OI is found to be the best because it accounts for differences in elevation and hence does not introduce spurious snow cover to lower elevations.

All model initializations are unable to consistently estimate the upscaled area-averaged SD in mountainous boxes. RAP has the best performance, while the other three (GFS, CFS, and NAM) show similarly poor results. However, even RAP's performance is not consistent across all water years, largely due to the model upgrades in February 2012 and February 2014. The RAP revision in February 2012 involved upgrades to the implementation of the WRF Model framework (McClung 2012), a larger model domain (covering all of North America), and upgrades to the data assimilation scheme that explain the break in data for early 2012. The revision in February 2014 involved changes to the snow data assimilation procedure to allow the IMS product to add snow where the background is snow-free and an adjustment of the treatment of snow albedo in the LSM (McClung 2014; Smirnova et al. 2014). However, the exact cause for the large drop in SD (Fig. 5) is unknown. The better performance of RAP than NAM (with a similar grid size) suggests that the use of LSM SD (rather than the direct replacement of LSM SD) should be emphasized in snow initialization.

Models utilizing the AFWA data underestimated SD the most out of all products tested. GFS and CFS (both

TABLE 4. December and April ratios (monthly mean of the ratio between daily model initialization and observations) of SD, ratios of SWE, and ρ for GFS, CFS, and NAM. The averages for WYs 2012–14 over the six boxes in the CONUS (i.e., excluding AKM and AKS) are also provided.

	GFS			CFS			NAM			SNODAS		
	SD	SWE	ρ	SD	SWE	ρ	SD	SWE	ρ	SD	SWE	ρ
December												
WI	1.05	—	—	1.05	—	—	1.02	—	—	1.10	—	—
AKM	0.85	0.70	0.82	0.90	0.74	0.82	—	—	—	—	—	—
CO	0.82	0.32	0.43	0.67	0.27	0.43	0.80	0.64	0.86	0.85	0.68	0.80
MT	0.45	0.18	0.41	0.40	0.16	0.41	0.47	0.38	0.83	1.11	1.02	0.93
YS	0.62	0.24	0.39	0.76	0.29	0.39	0.66	0.51	0.77	1.04	0.85	0.82
AKS	0.54	0.27	0.49	0.73	0.31	0.43	—	—	—	—	—	—
WA	0.28	0.09	0.29	0.26	0.08	0.29	0.29	0.20	0.64	1.13	0.85	0.76
ID	0.27	0.10	0.39	0.32	0.13	0.39	0.25	0.21	0.81	0.93	0.79	0.86
Avg without AKM and AKS	0.58	0.19	0.38	0.58	0.19	0.38	0.58	0.39	0.78	1.03	0.84	0.83
April												
WI	1.20	—	—	0.90	—	—	2.37	—	—	2.50	—	—
AKM	0.63	0.30	0.48	0.80	0.39	0.48	—	—	—	—	—	—
CO	0.11	0.03	0.26	0.09	0.02	0.26	0.09	0.05	0.56	0.81	0.62	0.78
MT	0.04	0.01	0.26	0.03	0.01	0.26	0.03	0.02	0.59	1.05	0.99	0.94
YS	0.17	0.04	0.26	0.14	0.04	0.26	0.13	0.08	0.58	0.95	0.78	0.83
AKS	0.21	0.06	0.28	0.24	0.06	0.25	—	—	—	—	—	—
WA	0.02	0.00	0.22	0.01	0.00	0.22	0.01	0.01	0.54	0.98	0.82	0.83
ID	0.04	0.01	0.25	0.03	0.01	0.25	0.03	0.02	0.60	0.77	0.67	0.87
Avg without AKM and AKS	0.26	0.02	0.25	0.20	0.02	0.25	0.44	0.04	0.57	1.18	0.78	0.85

at $0.5^\circ \times 0.5^\circ$ grids) as well as NAM (at $12\text{ km} \times 12\text{ km}$ grids) all utilize the AFWA SD data and are unable to reproduce the area-averaged SD created by upscaled observations. This conclusion is not much affected by the model resolution as the higher-resolution NAM performs similarly to the lower-resolution GFS with similar initialization procedures, while RAP (with a similar grid spacing as NAM) performs much better without the inclusion of AFWA data. The SD differences between GFS and CFS are partly explained by their different initialization procedures. While both models utilize the AFWA SD and IMS snow cover, GFS applies the AFWA SD directly. However, CFS compares the first-guess SDs to the AFWA values as explained in section 2c. All products are more deficient in SD and SWE over mountainous areas than over flat areas regardless of grid spacing. Initialization in low relief areas is considerably easier because of fewer environmental effects on snow distribution and amount. An important note is that the aforementioned deficiencies are not solely caused by data utilized in the AFWA product because of the ability of forecasters to alter SD, as mentioned in section 2c.

Similar to SD, the SWE initializations are also much lower than our SWE estimates based on observations (though we were unable to obtain SWE data for RAP, so we cannot test it). Furthermore, the application of constant density in GFS, CFS, and NAM exacerbates this underestimation. SWE ratios (between models and

observations) are lower than the corresponding SD ratios for GFS, CFS, and NAM by 38%–76% (ignoring WI, AKM, and AKS) in April. Decreases are lowest for NAM because of the higher constant snow density of 200 kg m^{-3} (compared to 100 kg m^{-3} for GFS and CFS). Ratios decrease from December to April in these models because the actual snow density increases as winter progresses. This application of spatially and temporally constant snow density is a particularly large problem in the NCEP model initializations, and we are currently creating a physically based snow density model for operational implementation.

Uncertainty quantification tests indicate that the above conclusions are insensitive to the number of sites used (with 50%, 75%, or 90% data denial) in the CO box or any one of the three combined methods (PWBIN regression for the first-guess fields; and IDW, KRIG, or OI method) used for upscaling in situ data to box averages. One limitation of our PWBIN–OI method is that a box with very few observations may not be representative of area-averaged SD or SWE. A combination of PWBIN–OI with additional information (e.g., remotely sensed datasets and LSM estimates) as discussed below would be necessary for remote locations.

The CMC daily global SD product has been widely used to evaluate models in the past. However, its performance is consistently worse than RAP. It performs similarly to GFS, CFS, and NAM with large underestimations of SD in the CONUS. This underestimation

may partly stem from observations utilized by CMC (SYNOP, METAR, and SA reports). While the exclusion of SNOTEL and COOP observations allows for an independent evaluation dataset, the performance of CMC may improve with inclusion of SNOTEL and COOP observations.

SNODAS assimilates all available snow measurements from ground, airborne, and satellite sources (including SNOTEL and COOP data), and hence it is not a completely independent product from our method. It agrees with our product very well. Compared with SNODAS, our method assimilates SNOTEL and COOP data only without using any LSMs. Furthermore, SNODAS is only available for the CONUS, which limits its ability to validate global models (or even over Alaska). Further examination of data from SNODAS, the Global Land Data Assimilation System (GLDAS), and the North American Land Data Assimilation System (NLDAS) will be performed in subsequent papers.

Because of the importance of snow initialization, future work is needed to develop a method for initializing SD and SWE in NCEP operational models and to evaluate the improvement through rerunning models with an updated initialization. Current attempts to assimilate remotely sensed snow quantities have successfully utilized ensemble Kalman filters (EnKFs; Liu et al. 2013; Kumar et al. 2014, 2015) to improve assimilated snow states. This approach may also be used to combine the strengths of all available snow datasets (e.g., PWBIN-OI, integrated LSM snow, and remotely sensed data) into a best estimate of global SD and SWE.

Furthermore, all forecast models are in a constant state of flux. An overview and full list of up-to-date changes for GFS and NAM are available online (<http://www.emc.ncep.noaa.gov>). Both GFS and NAM changed snow initialization procedures in WY 2015. Direct replacement of SD is no longer utilized, which acknowledges the value of LSM SD and SWE to evolve the snowpack. While GFS now follows a similar procedure as CFS, NAM dramatically changed snow initialization procedures with the removal of AFWA data completely. Instead, NAM cycles snow and updates snow cover extent with the 4-km IMS product. The evaluation of these new products will be a future task, but changes in GFS are not expected to significantly improve model performance because of the similar performance between GFS and CFS in this study.

Acknowledgments. This work was supported by the Idaho Power Company and NOAA (Award NA13NES4400003). Three anonymous reviewers and the Chief Editor (Dr. Christa D. Peters-Lidard) are thanked for their constructive and helpful comments and suggestions. NWS COOP data were downloaded from the National

Centers for Environmental Information (<http://www.ncdc.noaa.gov/snow-and-ice/daily-snow/>). The SNOTEL data were downloaded through the MesoWest data portal (Horel et al. 2002; <http://mesowest.utah.edu>). The source of the RAP model snow cycling material is available online (<http://meted.ucar.edu/>).

REFERENCES

- AFWA, 2013: Algorithm description document for the Air Force Weather Agency Snow Depth Analysis Model. 10 pp.
- Ault, T. W., K. P. Czajkowski, T. Benko, J. Coss, J. Struble, A. Spongberg, M. Templin, and C. Gross, 2006: Validation of the MODIS snow product and cloud mask using student and NWS cooperative station observations in the Lower Great Lakes region. *Remote Sens. Environ.*, **105**, 341–353, doi:10.1016/j.rse.2006.07.004.
- Balk, B., and K. Elder, 2000: Combining binary decision tree and geostatistical methods to estimate snow distribution in a mountain watershed. *Water Resour. Res.*, **36**, 13–26, doi:10.1029/1999WR900251.
- Barrett, A., 2003: National Operational Hydrologic Remote Sensing Center Snow Data Assimilation System (SNODAS) products at NSIDC. NSIDC Special Rep. 11, 19 pp. [Available online at https://nsidc.org/pubs/documents/special/nsidc_special_report_11.pdf.]
- Baxter, M. A., C. E. Graves, and J. T. Moore, 2005: A climatology of snow-to-liquid ratio for the contiguous United States. *Wea. Forecasting*, **20**, 729–744, doi:10.1175/WAF856.1.
- Benjamin, S. G., and Coauthors, 2004: An hourly assimilation-forecast cycle: The RUC. *Mon. Wea. Rev.*, **132**, 495–518, doi:10.1175/1520-0493(2004)132<0495:AHACTR>2.0.CO;2.
- Brasnett, B., 1999: A global analysis for snow depth for numerical weather prediction. *J. Appl. Meteor.*, **38**, 726–740, doi:10.1175/1520-0450(1999)038<0726:AGAOSD>2.0.CO;2.
- Brown, R., and B. Brasnett, 2010: Canadian Meteorological Centre (CMC) daily snow depth analysis data, version 1 (updated annually). National Snow and Ice Data Center, accessed April 2015, doi:10.5067/W9FOYWH0EQZ3.
- , —, and D. Robinson, 2001: Development of a gridded North American monthly snow depth and snow water equivalent dataset for GCM validation. *Proc. 58th Eastern Snow Conf.*, Ottawa, Ontario, Canada, Eastern Snow Conference, 215–217. [Available online at http://www.easternsnow.org/proceedings/2002/018_Brown.pdf.]
- Christensen, N. S., and D. P. Lettenmaier, 2007: A multimodel ensemble approach to assessment of climate change impacts on the hydrology and water resources of the Colorado River basin. *Hydrol. Earth Syst. Sci.*, **11**, 1417–1434, doi:10.5194/hess-11-1417-2007.
- COMET, 2015: Operational Models Matrix: Characteristics of NWP and related forecast models. MetEd, accessed 1 February 2015. [Available online at <http://www.meted.ucar.edu/nwp/pcu2/>.]
- Dietz, A. J., C. Kuenzer, U. Gessner, and S. Dech, 2012: Remote sensing of snow—A review of available methods. *Int. J. Remote Sens.*, **33**, 4094–4134, doi:10.1080/01431161.2011.640964.
- Dixon, D., S. Boon, and U. Silins, 2014: Watershed-scale controls on snow accumulation in a small montane watershed, southwestern Alberta, Canada. *Hydrol. Processes*, **28**, 1294–1306, doi:10.1002/hyp.9667.
- Drusch, M., D. Vasiljevic, and P. Viterbo, 2004: ECMWF's global snow analysis: Assessment and revision based on satellite

- observations. *J. Appl. Meteor.*, **43**, 1282–1294, doi:[10.1175/1520-0450\(2004\)043<1282:EGSAAA>2.0.CO;2](https://doi.org/10.1175/1520-0450(2004)043<1282:EGSAAA>2.0.CO;2).
- EMC, 2003: The GFS Atmospheric Model. NOAA/NWS/NCEP, NCEP Office Note 442, 14 pp. [Available online at <http://www.lib.ncep.noaa.gov/ncepofficenotes/files/on442.pdf>.]
- , 2015: Mesoscale modeling branch FAQ. NOAA/NWS, accessed 19 May 2016. [Available online at www.emc.ncep.noaa.gov/NAM/faq.php.]
- Erxleben, J., K. Elder, and R. Davis, 2002: Comparison of spatial interpolation methods for estimating snow distribution in the Colorado Rocky Mountains. *Hydrol. Processes*, **16**, 3627–3649, doi:[10.1002/hyp.1239](https://doi.org/10.1002/hyp.1239).
- ESRL, 2015: Diagnostic output fields for the Rapid Refresh and HRRR. NOAA/ESRL, accessed 19 May 2016. [Available online at http://ruc.noaa.gov/rr/RAP_var_diagnosis.html.]
- Farr, T. G., and Coauthors, 2007: The Shuttle Radar Topography Mission. *Rev. Geophys.*, **45**, RG2004, doi:[10.1029/2005RG000183](https://doi.org/10.1029/2005RG000183).
- Fassnacht, S. R., K. A. Dressler, and R. C. Bales, 2003: Snow water equivalent interpolation for the Colorado River basin from snow telemetry (SNOTEL) data. *Water Resour. Res.*, **39**, 1208, doi:[10.1029/2002WR001512](https://doi.org/10.1029/2002WR001512).
- Foster, D. J., Jr., and R. D. Davy, 1988: Global snow depth climatology. Rep. USAFETAC/TN-88/006, 50 pp. [Available online at <http://www.dtic.mil/dtic/tr/fulltext/u2/a203969.pdf>.]
- Frei, A., M. Tedesco, S. Lee, J. Foster, D. K. Hall, R. Kelly, and D. A. Robinson, 2012: A Review of global satellite-derived snow products. *Adv. Space Res.*, **50**, 1007–1029, doi:[10.1016/j.asr.2011.12.021](https://doi.org/10.1016/j.asr.2011.12.021).
- Grunewald, T., and Coauthors, 2013: Statistical modelling of the snow depth distribution in open alpine terrain. *Hydrol. Earth Syst. Sci.*, **17**, 3005–3021, doi:[10.5194/hess-17-3005-2013](https://doi.org/10.5194/hess-17-3005-2013).
- Helfrich, S. R., D. McNamara, B. H. Ramsay, T. Baldwin, and T. Kasheta, 2007: Enhancements to, and forthcoming developments in the Interactive Multisensor Snow and Ice Mapping System (IMS). *Hydrol. Processes*, **21**, 1576–1588, doi:[10.1002/hyp.6720](https://doi.org/10.1002/hyp.6720).
- Horel, J., and Coauthors, 2002: MesoWest: Cooperative Mesonets in the Western United States. *Bull. Amer. Meteor. Soc.*, **83**, 211–225, doi:[10.1175/1520-0477\(2002\)083<0211:MCMITW>2.3.CO;2](https://doi.org/10.1175/1520-0477(2002)083<0211:MCMITW>2.3.CO;2).
- James, T., A. Evans, E. Madly, and C. Kelly, 2014: The economic importance of the Colorado River to the basin region. Final Rep., L. William Seidman Research Institute, Arizona State University, 54 pp. [Available online at <http://www.protectflows.com/wp-content/uploads/2015/01/PTF-Final-121814.pdf>.]
- Janjić, Z., and R. L. Gall, 2012: Scientific documentation of the NCEP nonhydrostatic multiscale model on the B grid (NMMB). Part 1 Dynamics. NCAR Tech. Note NCAR/TN-489+STR, 75 pp., doi:[10.5065/D6WH2MZX](https://doi.org/10.5065/D6WH2MZX).
- Jet Propulsion Laboratory, 2003: SRTM30 documentation. California Institute of Technology, 9 pp. [Available online at http://dds.cr.usgs.gov/srtm/version2_1/SRTM30/srtm30_documentation.pdf.]
- Kirchner, P. B., R. C. Bales, N. P. Molotch, J. Flanagan, and Q. Guo, 2014: LiDAR measurement of seasonal snow accumulation along an elevation gradient in the southern Sierra Nevada, California. *Hydrol. Earth Syst. Sci.*, **18**, 4261–4275, doi:[10.5194/hess-18-4261-2014](https://doi.org/10.5194/hess-18-4261-2014).
- Kumar, S., and Coauthors, 2014: Assimilation of remotely sensed soil moisture and snow depth retrievals for drought estimation. *J. Hydrometeorol.*, **15**, 2446–2469, doi:[10.1175/JHM-D-13-0132.1](https://doi.org/10.1175/JHM-D-13-0132.1).
- , C. D. Peters-Lidar, K. R. Arsenault, A. Getirana, D. Mocko, and Y. Liu, 2015: Quantifying the added value of snow cover area observations in passive microwave snow depth data assimilation. *J. Hydrometeorol.*, **16**, 1736–1741, doi:[10.1175/JHM-D-15-0021.1](https://doi.org/10.1175/JHM-D-15-0021.1).
- Liu, Y., C. D. Peters-Lidar, S. Kumar, J. L. Foster, M. Shaw, Y. Tian, and G. M. Fall, 2013: Assimilating satellite-based snow depth and snow cover products for improving snow predictions in Alaska. *Adv. Water Resour.*, **54**, 208–227, doi:[10.1016/j.advwatres.2013.02.005](https://doi.org/10.1016/j.advwatres.2013.02.005).
- López-Moreno, J. I., and D. Nogués-Bravo, 2006: Interpolating local snow depth data: An evaluation of methods. *Hydrol. Processes*, **20**, 2217–2232, doi:[10.1002/hyp.6199](https://doi.org/10.1002/hyp.6199).
- McClung, T., 2012: Technical Implementation Notice 11-53, amended. TIN11-53, National Weather Service. [Available online at http://www.nws.noaa.gov/os/notification/tin11-53ruc_rapaac.txt.]
- , 2014: Technical Implementation Notice 13-38, amended. TIN13-38, National Weather Service. [Available online at http://www.nws.noaa.gov/om/notification/tin13-38rap_aab.htm.]
- Meromy, L., N. P. Molotch, T. E. Link, S. R. Fassnacht, and R. Rice, 2013: Subgrid variability of snow water equivalent at operational snow stations in the western USA. *Hydrol. Processes*, **27**, 2383–2400, doi:[10.1002/hyp.9355](https://doi.org/10.1002/hyp.9355).
- Mizukami, N., and S. Perica, 2012: Towards improved snow water equivalent retrieval algorithms for satellite passive microwave data over the mountainous basins of western USA. *Hydrol. Processes*, **26**, 1991–2002, doi:[10.1002/hyp.8333](https://doi.org/10.1002/hyp.8333).
- Molotch, N. P., and R. C. Bales, 2005: Scaling snow observations from the point to the grid element: Implications for observation network design. *Water Resour. Res.*, **41**, W11421, doi:[10.1029/2005WR004229](https://doi.org/10.1029/2005WR004229).
- 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, doi:[10.1029/2007JD008674](https://doi.org/10.1029/2007JD008674).
- Northrop Grumman, 2002: Algorithm and Data User Manual (ADUM) for the Special Sensor Microwave Imager/Sounder (SSMIS). Tech. Rep. 12621, 65 pp. [Available online at <https://www.ncdc.noaa.gov/oa/rsad/ssmi/swath/adum-ssmis-description.pdf>.]
- Palmer, P. L., 1988: The SCS Snow Survey Water Supply Forecasting Program: Current operations and future directions. *Proc. 56th Annual Western Snow Conf.*, Kalispell, MT, Western Snow Conference, 43–51. [Available online at <http://westernsnowconference.org/sites/westernsnowconference.org/PDFs/1988Palmer.pdf>.]
- 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, doi:[10.1029/2003JD003994](https://doi.org/10.1029/2003JD003994).
- Reichle, R. H., R. D. Koster, G. J. M. Lannoy, B. A. Forman, Q. Liu, and S. P. P. Mahanama, 2011: Assessment and enhancement of MERRA land surface hydrology estimates. *J. Climate*, **24**, 6322–6338, doi:[10.1175/JCLI-D-10-05033.1](https://doi.org/10.1175/JCLI-D-10-05033.1).
- Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1057, doi:[10.1175/2010BAMS3001.1](https://doi.org/10.1175/2010BAMS3001.1).
- , and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, doi:[10.1175/JCLI-D-12-00823.1](https://doi.org/10.1175/JCLI-D-12-00823.1).
- Serreze, M. C., M. P. Clark, and R. L. Armstrong, 1999: Characteristics of the western United States snowpack from

- Snowpack Telemetry (SNOTEL) data. *Water Resour. Res.*, **35**, 2145–2160, doi:[10.1029/1999WR900090](https://doi.org/10.1029/1999WR900090).
- Slater, A. G., and M. P. Clark, 2006: Snow data assimilation via an ensemble Kalman filter. *J. Hydrometeor.*, **7**, 478–493, doi:[10.1175/JHM505.1](https://doi.org/10.1175/JHM505.1).
- Smirnova, T. G., J. M. Brown, and S. Benjamin, 2014: Recent Developments in RUC Land Surface Model (RUC LSM) Implemented in Operational Rapid Refresh (RAP) at NCEP. *26th Conf. on Weather Analysis and Forecasting/22nd Conf. on Numerical Weather Prediction*, Atlanta, GA, Amer. Meteor. Soc., 12.2. [Available online at <https://ams.confex.com/ams/94Annual/webprogram/Paper234690.html>.]
- Wi, S., F. Dominguez, M. Durcik, J. Valdes, H. F. Diaz, and C. L. Castro, 2012: Climate change projection of snowfall in the Colorado River basin using dynamic downscaling. *Water Resour. Res.*, **48**, W05504, doi:[10.1029/2011WR010674](https://doi.org/10.1029/2011WR010674).
- Yang, Z.-L., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *J. Geophys. Res.*, **116**, D12110, doi:[10.1029/2010JD015140](https://doi.org/10.1029/2010JD015140).