Evaluation of Recent NCEP Operational Model Upgrades for Cool-Season Precipitation Forecasting over the Western Conterminous United States

MARCEL CARON AND W. JAMES STEENBURGH

Department of Atmospheric Sciences, University of Utah, Salt Lake City, Utah

(Manuscript received 2 September 2019, in final form 4 March 2020)

ABSTRACT

In August 2018 and June 2019, NCEP upgraded the operational versions of the High-Resolution Rapid Refresh (HRRR) and Global Forecast System (GFS), respectively. To inform forecasters and model developers about changes in the capabilities and biases of these modeling systems over the western conterminous United States (CONUS), we validate and compare precipitation forecasts produced by the experimental, preoperational HRRRv3 and GFSv15.0 with the then operational HRRRv2 and GFSv14 during the 2017/18 October–March cool season. We also compare the GFSv14 and GFSv15.0 with the operational, high-resolution configuration of the ECMWF Integrated Forecasting System (HRES). We validate using observations from Automated Surface and Weather Observing System (ASOS/AWOS) stations, which are located primarily in the lowlands, and observations from Snowpack Telemetry (SNOTEL) stations, which are located primarily in the uplands. Changes in bias and skill from HRRRv2 to HRRRv3 are small, with HRRRv3 exhibiting slightly higher (but statistically indistinguishable at a 95% confidence level) equitable threat scores. The GFSv14, GFSv15.0, and HRES all exhibit a wet bias at lower elevations and neutral or dry bias at upper elevations, reflecting insufficient terrain representation. GFSv15.0 performance is comparable to GFSv14 at day 1 and superior at day 3, but lags HRES. These results establish a baseline for current operational HRRR and GFS precipitation capabilities over the western CONUS and are consistent with steady or improving NCEP model performance.

1. Introduction

Upgrades to operational forecast systems introduce challenges for both operational meteorologists and model developers. Operational meteorologists rely on knowledge of model biases and prior performance to make reliable weather forecasts and assess potential societal impacts. Model developers require knowledge of model capabilities to address model deficiencies and advance model performance. Since 2018, NCEP has upgraded two major operational forecast systems: the High-Resolution Rapid Refresh (HRRR) and the Global Forecast System (GFS). The HRRR operates at 3-km grid spacing and provides short-range forecasts for the conterminous United States (CONUS). The GFS operates at an effective grid spacing of 13 km and provides short- to medium-range global forecasts. Both modeling systems contribute to the National Blend of Models (NBM), which heavily informs NWS forecasts (Craven et al. 2018).

Although model validation is a component of the development and upgrade cycle at NCEP, it does not include detailed validation of regional precipitation forecasts. Of concern for this paper are cool-season (October–March) precipitation events over the western CONUS, which are strongly influenced by the interaction of synoptic systems with orography and often produce snow, posing critical challenges for transportation and public safety (Andrey et al. 2001; Birkeland and Mock 2001; Seeherman and Liu 2015). Atmospheric rivers and other landfalling, extratropical disturbances contribute a substantial fraction of total cool-season precipitation over the region (Rutz et al. 2014; Barbero et al. 2019), with mean precipitation generally increasing with elevation (Daly et al. 1994). Nevertheless, individual storm periods can feature precipitation–altitude relationships that depart from that expected from climatology, presenting a challenge for operational and numerical weather prediction (Steenburgh 2003; James and Houze 2005; Minder et al. 2008). Forecast skill also decreases from the Pacific coast to the western interior, even for relatively high-resolution forecast systems

Corresponding author: W. James Steenburgh, jim.steenburgh@utah.edu

DOI: 10.1175/WAF-D-19-0182.1

(Lewis et al. 2017; Gowan et al. 2018). This decrease may reflect the finer-scale nature of the topography and the reduced spatial coherence of cool-season precipitation events downstream of the Cascade–Sierra Ranges (Serreze et al. 2001; Parker and Abatzoglou 2016; Touma et al. 2018).

Recent studies indicate that model resolution contributes to spatial variations in precipitation bias and skill among forecast systems over the western United States (Gowan et al. 2018). Forecast systems that feature smooth orography and fail to resolve terrain complexity sometimes produce excessive lowland and insufficient upland precipitation. Downscaling can partially address this deficiency (Lewis et al. 2017). Higher-resolution convection-allowing models like the HRRR better resolve regional terrain features and produce improved skill as measured by traditional skill scores (Gowan et al. 2018). Nevertheless, errors at high resolution evolve more rapidly in time and can contribute to deterioration in forecast skill at short lead times (Lorenz 1969; Prein et al. 2015; Clark et al. 2016).

In this paper we examine the performance of the experimental, preoperational HRRRv3 and GFSv15.0 compared to their predecessor operational versions, HRRRv2 and GFSv14, respectively. The HRRRv3 upgrades include an improved planetary boundary layer [Mellor–Yamada–Nakanishi–Niino (MYNN); Nakanishi and Niino (2009)] and a new, hybrid vertical coordinate (Simmons and Strüfing 1983; Collins et al. 2004). The GFSv15.0 features a new finite volume cubed-sphere dynamical core (Chen et al. 2018; Hazelton et al. 2018) and includes the GFDL six-category bulk cloud microphysics scheme (described in Chen and Lin 2013). We specifically evaluate cool-season precipitation forecasts over the western CONUS, at both lowland and upland locations, to identify modeling system capabilities and biases for forecasters and model developers, as well as establish a baseline of current NCEP operational model performance.

The remainder of this paper is organized as follows. Section 2 describes the models and observational data used for the evaluation, as well as the validation methodology. Section 3 examines and describes the results and performance of the experimental modeling systems relative to their operational predecessors and compares GFS performance to the operational, high-resolution configuration of the ECMWF Integrated Forecasting System (HRES). A summary of the results follows in section 4.

2. Data and methods

a. Forecast systems

The HRRR is an hourly updating forecast system that is nested within the 13-km Rapid Refresh (RAP) and provides forecasts for the CONUS at 3-km grid spacing (Benjamin et al. 2016; Myrick 2018). During the 2017/18 cool season, which is the focus of this study, NCEP produced operational forecasts with HRRRv2, whereas the NOAA/Earth System Research Laboratory (ESRL) ran the experimental HRRRv3. HRRRv2 uses the Advanced Research version of WRF, version 3.6, with physics packages and assimilation procedures described in Benjamin et al. (2016). HRRRv3 uses the Advanced Research version of WRF version 3.8, with model physics, numerics, assimilated datasets, and assimilation techniques described by NOAA (2018). HRRRv2 forecasts were obtained from the NCEP Operational Model Archive and Distribution System (NOMADS), whereas HRRRv3 forecasts were provided by ESRL. The HRRRv3 became operational at NCEP in August 2018.

The GFS is a global forecast system. During the 2017/18 cool season, NCEP produced operational forecasts using GFSv14, a global spectral model with T1534 horizontal resolution (~13 km) for the initial 10-day forecast period. Major GFS parameterization and data assimilation techniques are described in McClung (2014), NWS (2016), and Myrick (2017). The GFSv15.0 represents a major upgrade and uses a finite volume cubed-sphere dynamical core developed at GFDL with an effective horizontal resolution comparable to GFSv14. Physics packages are based on GFSv14, except for the replacement of the Zhao–Carr microphysics scheme with the GFDL microphysics scheme (Yang 2018), updates or new parameterizations for ozone and water vapor photochemistry, and a revised bare-soil evaporation scheme (Tallapragada and Yang 2018). Operational GFSv14 forecasts and GFSv15.0 reforecasts were obtained from the NCEP Environmental Modeling Center. Ultimately, the operational GFS was upgraded from GFSv14 to GFSv15.1 in June 2019 rather than GFS15.0, with GFSv15.1 including some improvements that reduce but do not eliminate a near-surface cold bias that led to excessive accumulated snow. However, we focus on liquid precipitation equivalent and tests indicate that GFSv15.0 and GFSv15.1 produce relatively similar quantitative precipitation forecasts (A. Bentley, NCEP, 2019, personal communication).

We also compare GFSv14 and GFSv15.0 forecasts with HRES, a global forecast model developed and run by ECMWF. During the 2017/18 cool season, HRES was based on the ECMWF Integrated Forecasting System (IFS) cycle 43r3 with a 0.07° effective horizontal resolution over an octahedral reduced Gaussian grid (ECMWF 2019). Parameterizations are described by Roberts et al. (2018). Operational HRES forecasts were provided by ECMWF.

b. Precipitation observations

Precipitation validation focuses on the CONUS west of 102.5°W and uses observations from the Automated

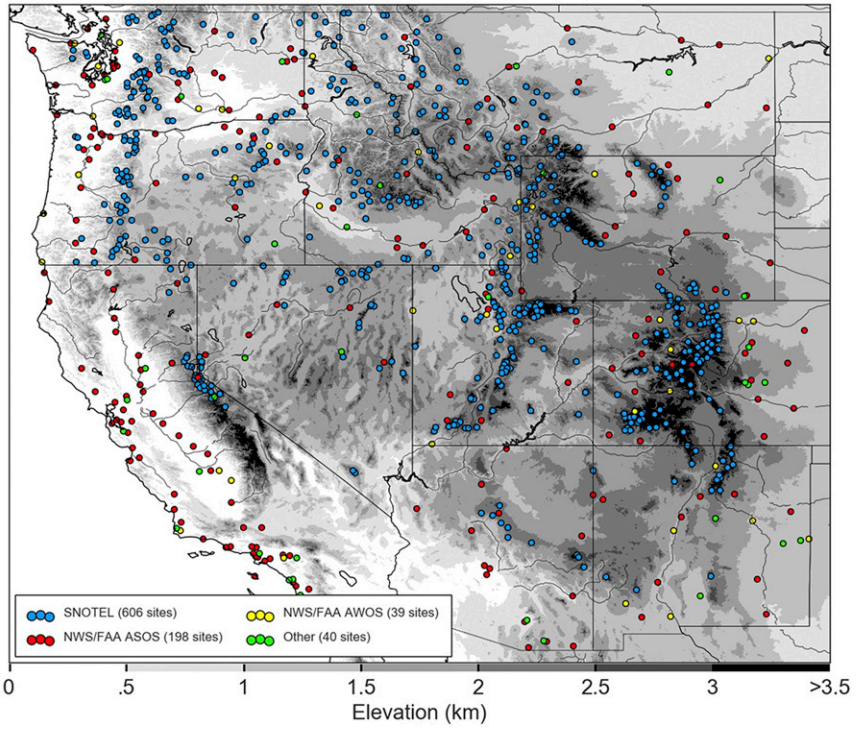

FIG. 1. ASOS/AWOS (red) and SNOTEL (blue) stations used for this study with 30-arc-s topography (km MSL; shaded). Station classification based on FAA (2020).

Surface and Weather Observing System network operated by the NWS, Federal Aviation Administration (FAA), and Department of Defense (DoD), and the Snowpack Telemetry network operated by the Natural Resources Conservation Service (hereafter SNOTEL) (Fig. 1). The former consists primarily of Automated Surface Observing System (ASOS) and Automated Weather Observing System (AWOS) units and is hereafter referred to as ASOS/AWOS. ASOS stations are maintained by NWS electronics technicians and measure precipitation in 0.01-in. (0.254 mm) increments using either a standard heated tipping bucket with a vinyl alter-style wind shield or an all-weather precipitation accumulation gauge with a Tretyakov wind shield (Martinaitis et al. 2015). The standard heated tipping buckets are implemented at a majority of ASOS stations, but the all-weather precipitation accumulation gauge has been installed at some stations since 2003 (NWS 2009; Martinaitis et al. 2015). AWOS stations are maintained by the FAA or local airport boards and use tipping buckets or all-weather precipitation accumulation gauges that may be unshielded (FAA 2017). Additionally, we include automated precipitation observations from stations maintained primarily by the Department of Defense and classified as other in Fig. 1.

ASOS/AWOS precipitation gauge undercatch of snowfall increases with wind speed and can result in the underreporting of liquid precipitation equivalent as large as 20%–50% during snowfall (Greeney et al. 2005; Rasmussen et al. 2012; Martinaitis et al. 2015). Additionally, losses due to evaporation or sublimation can occur with the heated tipping bucket and snow can stick to the orifice or sidewalls of the all-weather precipitation gauge, resulting in a delay in snowfall measurement (Martinaitis et al. 2015).

ASOS/AWOS data were obtained from Synoptic Data, a Public Benefit Corporation owned in part by the University of Utah, using their Application Program Interface (https://synopticlabs.org/synoptic/) and were quality controlled following procedures described by Horel et al. (2002) and in documentation available from Synoptic Data. To reduce sampling issues, stations were chosen that recorded five or more days with measurable precipitation [i.e., ≥0.01 in. (0.254 mm); Durre et al. 2013] and received ≥0.5 in. (12.7 mm) of total accumulated precipitation on days for which model forecasts were available. The resulting 277 stations (Fig. 1)—situated predominantly (but not exclusively) in lowland areas and located mainly at airports—provided 6-h accumulated precipitation observations, which were aggregated into 24-h totals. The 277 stations include 198 (out of a possible

224) NWS/FAA ASOS stations, 39 (out of a possible 278) NWS/FAA AWOS stations, and 40 additional (primarily DoD) stations. Questions about the quality of AWOS precipitation observations have been raised by operational meteorologists and the elimination of a large fraction of AWOS stations is consistent with those concerns. Results were not strongly sensitive to the inclusion of the 39 AWOS stations.

The density of utilized AWOS/ASOS stations is variable and greatest in the lowlands of western California, Oregon, and Washington, but in southeast California, southwest Arizona, and southern Nevada none of the available AWOS/ASOS stations observe sufficient precipitation or days with measurable precipitation to be included in the study.

SNOTEL stations are located at remote, sheltered, upland locations. Accumulated precipitation is measured hourly in 0.1-in. (2.54-mm) increments using a large-storage gauge. SNOTEL precipitation measurements exhibit an artificially driven diurnal cycle due to expansion and contraction of fluid in the gauge (USDA 2014). We limit this effect by using only 24-h accumulated precipitation measurements. Other errors are addressed by quality controlling data according to the methods described by Lewis et al. (2017), yielding data from 606 SNOTEL stations. Like ASOS/AWOS stations, undercatch remains a likely source of error for SNOTEL stations, although undercatch is typically smaller (10%–15%) during snowfall than ASOS/AWOS stations due to siting in wind-protected clearings (Ikeda et al. 2010). Similarly, delays in measurement can also occur due to snow sticking to the orifice or gauge walls. Thus, at ASOS/AWOS and SNOTEL stations that receive snowfall, it is likely that these sources of measurement error artificially shift model precipitation biases higher and reduce our ability to determine model accuracy.

Additional networks provide precipitation data for the western CONUS, such as the Remote Automated Weather Stations (RAWS) or the NWS COOP network. Although these networks do provide some high-quality observations, they are more heterogeneous and exhibit varying quality. For example, some networks employ unheated tipping buckets that misreport solid precipitation (e.g., RAWS). Daly et al. (2007) and Hashimoto et al. (2019) describe some of the measurement errors and quality control issues associated with COOP data. While likely valuable for future studies, data from these networks was not included here due to the time needed to develop more robust quality control techniques.

c. Validation

We validate model forecasts initialized between 0000 UTC 1 October 2017 and 1800 UTC 31 March 2018. The selection of the 2017/18 cool season reflects the availability of forecasts from all five modeling systems. To enable validation of 24-h precipitation (hereafter daily precipitation) using HRRRv2 and HRRRv3 forecasts, since the former only extends to 18 h, we combine the 6–18-h precipitation forecasts from the 0600 UTC and 1800 UTC initialized forecasts. GFSv14, GFSv15.0, and HRES validation focuses on 12–36-h (hereafter day 1) and 60–84-h (hereafter day 3) forecasts initialized at 0000 UTC (for brevity, intermediate statistics for day 2 are omitted). Periods when one or more model forecasts were missing were not included, resulting in validation of 112 HRRRv2/HRRRv3 and 115 GFSv14/GFSv15.0/HRES daily forecasts. This represents 62% and 63% of the 182 cool-season days, respectively. To compare modeled with observed precipitation, precipitation forecasts are bilinearly interpolated to each station location.

TABLE 1. Contingency table used for validation.

	Observed	
Forecast	Yes	No
Yes	Hit (a)	False alarm (b)
No	Miss (c)	Correct rejection (d)

Bias ratio is the ratio of forecast to observed precipitation integrated over the study period on days when forecasts are available. Means are calculated using all stations in each network. Voronoi-weighted (Weller et al. 2009) and unweighted methods to calculate the areal average bias ratios yielded statistically indiscernible results using a two-proportion Z test, so figures display only unweighted areal averages for simplicity. Other validation metrics use daily precipitation, the occurrence of which is sometimes referred to as an event. Frequency bias, for example, is the ratio of the number of forecast and observed daily precipitation events in a given size bin.

Additional measures employed to evaluate daily precipitation forecasts include the hit rate, false alarm ratio, and equitable threat score, which are based on a 2×2 contingency table (Table 1). As summarized in Mason (2003), hit rate (HR) is defined as

$$\mathrm{HR} = \frac{a}{a+c}, \tag{1}$$

false alarm ratio (FAR) is defined as

$$\mathrm{FAR} = \frac{b}{a+b}, \tag{2}$$

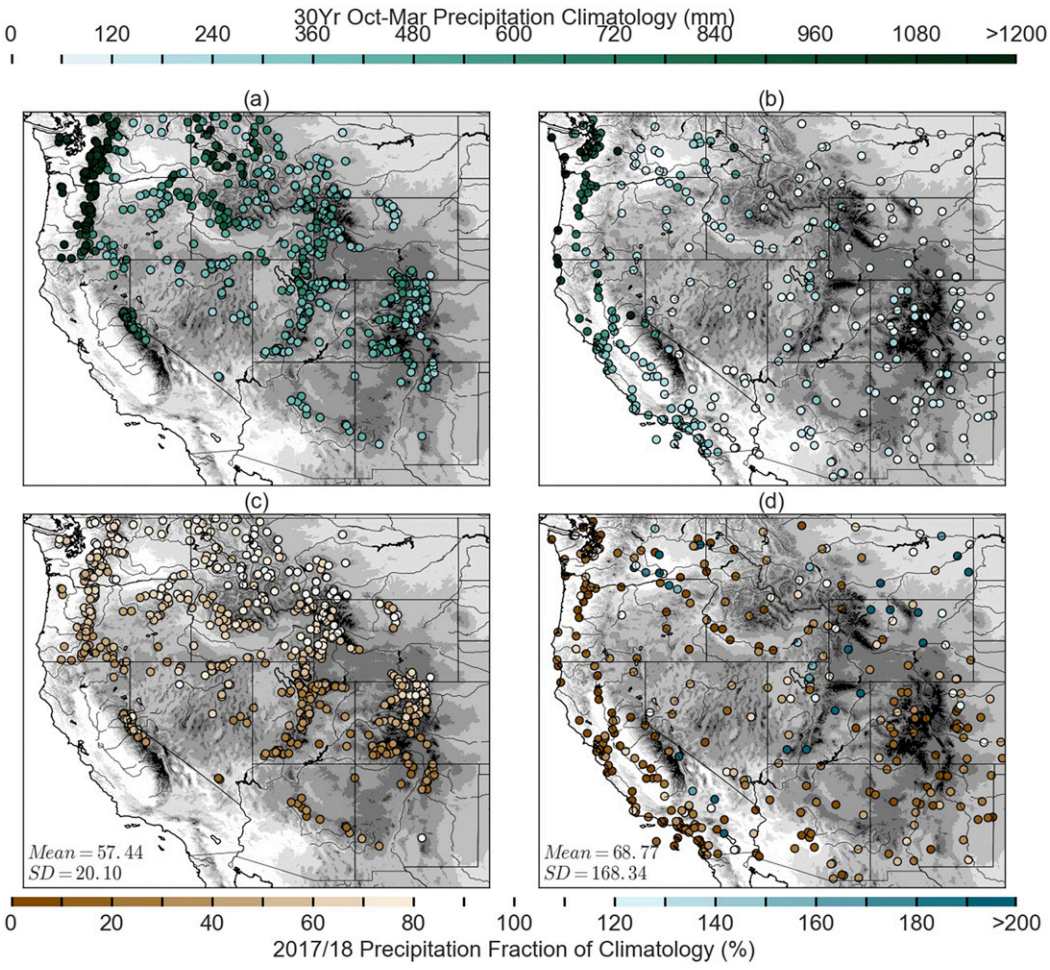

FIG. 2. The 30-yr average accumulated cool-season precipitation at (a) SNOTEL and (b) ASOS/AWOS stations [based on PRISM gridded climate data (Daly et al. 1994)], and 2017/18 cool-season total precipitation as a fraction of PRISM climatology at (c) SNOTEL and (d) ASOS/AWOS stations.

and equitable threat score (ETS) as

$$\mathrm{ETS} = \frac{a - a_{\mathrm{ref}}}{a - a_{\mathrm{ref}} + b + c}, \tag{3}$$

where

$$a_{\mathrm{ref}} = \frac{(a + c)(a + b)}{n}. \tag{4}$$

These measures are calculated using absolute precipitation amounts and percentile thresholds, the latter defined relative to the amount distribution for each model on all validation days, including those without measurable precipitation. We evaluate these measures using absolute precipitation thresholds and percentile thresholds based on 2017/18 cool-season precipitation events. The latter reduces the effects of model bias in the evaluation of the spatial accuracy of model forecasts (Roberts and Lean 2008; Mittermaier and Roberts 2010; Dey et al. 2014; Gowan et al. 2018).

3. Results

a. *Synopsis of 2017/18 cool-season precipitation*

The 30-yr (1981–2010) average October–March cool-season precipitation exhibits a strong dependence of precipitation on altitude across the western United States (Daly et al. 1994). For the SNOTEL stations used in this study, cool-season precipitation is greatest at stations in the Coastal, Cascade, and Olympic Mountains of the Pacific Northwest and locations in the northwest interior (Fig. 2a). For the ASOS/AWOS stations used in this study, cool-season precipitation is greatest along and near the Pacific coast of northern California, Oregon, and Washington and lower

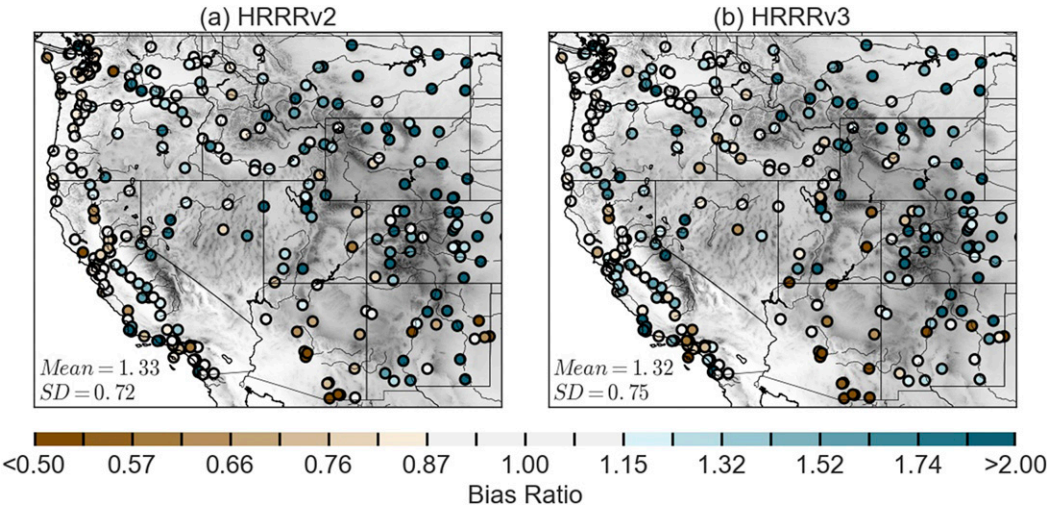

FIG. 3. (a) HRRRv2 and (b) HRRRv3 bias ratios at ASOS/AWOS stations with 30-arc-s topography (as in Fig. 1). Mean and standard deviation (SD) annotated.

in the valleys and basins of southern California and the western interior east of the Cascade–Sierra crest (Fig. 2b).

Integrated across all ASOS/AWOS and SNOTEL stations, the 2017/18 cool-season precipitation was about 40% below average. SNOTEL stations in the far north received near or slightly above-average precipitation, whereas stations farther south received below-average precipitation (Fig. 2c). This spatial pattern was comparatively less distinct at ASOS/AWOS stations, which exhibited less coherent regional patterns relative to average, especially east of the Cascade–Sierra crest (Fig. 2d). This likely reflects the relatively low frequency and spatial coherence of precipitation events east of the Cascade–Sierra crest (Rutz et al. 2014; Touma et al. 2018), which leads to undersampling at low elevation stations.

b. HRRR

During the 2017/18 cool season, the mean HRRRv2 bias ratio was 1.33 at ASOS/AWOS stations, indicating an overall wet bias (Fig. 3a). However, the bias ratio varied considerably from station to station, with a standard deviation of 0.72. Forecasts for stations in northern California, Oregon, and Washington west of the Cascade–Sierra crest exhibited primarily near-neutral or dry biases, whereas forecasts for stations east of the Cascade–Sierra crest predominantly exhibited near-neutral or wet biases. The HRRRv3 produced a similar mean bias ratio and standard deviation of 1.32

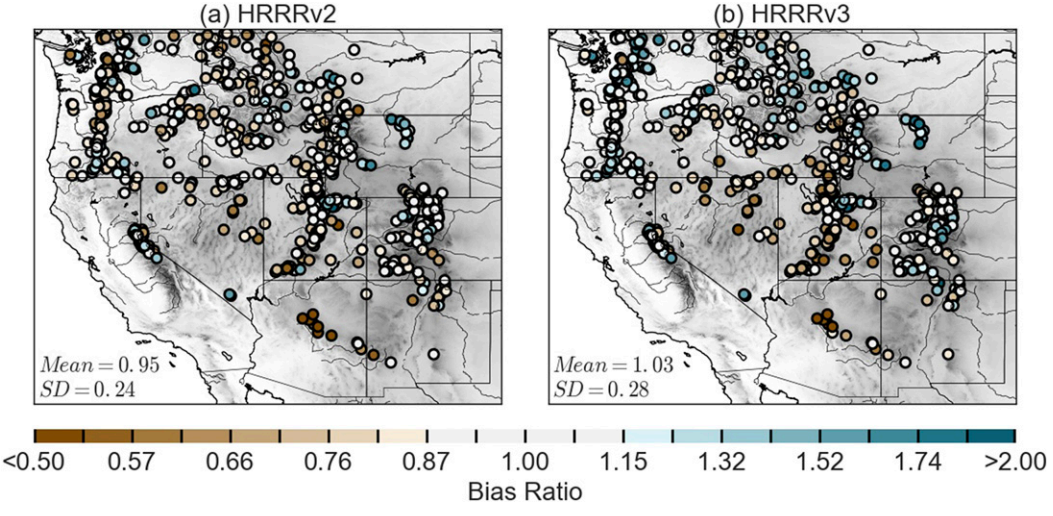

FIG. 4. As in Fig. 3, but for SNOTEL stations.

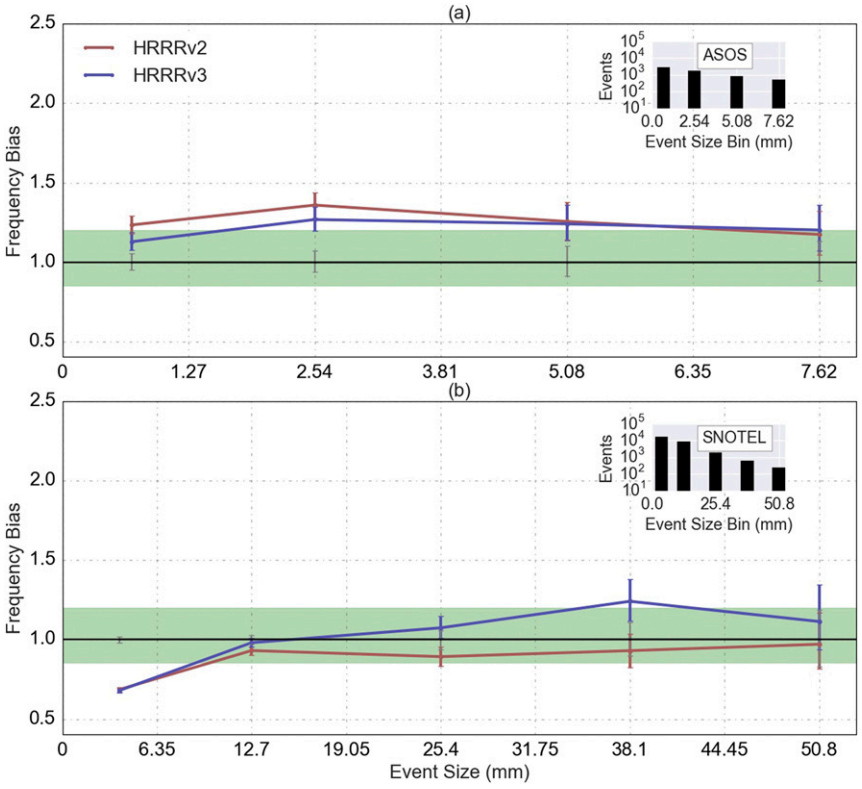

FIG. 5. HRRRv2 (red lines) and HRRRv3 (blue lines) frequency bias as a function of event size at (a) ASOS/AWOS and (b) SNOTEL stations. Number of events sampled into each bin shown in inset histograms. Green band shows 0.85–1.20 range defined as near neutral by the authors. Whiskers display 95% confidence intervals as determined using bootstrap resampling.

and 0.75, respectively, with a comparable spatial pattern of dry and wet biases at individual stations (Fig. 3b). At SNOTEL stations, the mean HRRRv2 bias ratio was 0.95, with greater consistency from station to station reflected in a low standard deviation (compared to forecasts for ASOS/AWOS stations) of 0.24 (Fig. 4a). Regions with larger dry (wet) biases include the Mogollon Rim of Arizona and ranges of eastern Nevada (Bighorn Mountains of Wyoming). The HRRRv3 was slightly wetter with a mean bias ratio of 1.03 and a small increase in standard deviation to 0.28 (Fig. 4b).

Frequency bias is the ratio of forecast to observed event frequency as a function of the observed event size (Fig. 5a). For convenience and following Lewis et al. (2017), we refer to a frequency bias of 0.85–1.20 as "near neutral" given the uncertainties in precipitation measurement. At ASOS/AWOS stations, we present frequency bias for events in four bins defined by lower and upper bounds [0.127–1.27 mm (0.005–0.05 in.), 1.27–3.81 mm (0.05–0.15 in.), 3.81–6.35 mm (0.15–0.25 in.), and 6.35–8.89 mm (0.25–0.35 in.)], represented in each graph by a central value. The lower bound is exclusive and the upper bound inclusive for all but the lowest bin [0.127–1.27 mm (0.005–0.05 in.)], for which we use model precipitation values $\geq$ 0.127 mm (0.005 in.) and observed precipitation values $\geq$ 0.254 mm (0.01 in.). Events $>$ 8.89 mm (0.35 in.) are not presented due to the small sample size. HRRRv2 exhibited frequency biases $>$ 1 at all event sizes and weak overprediction (i.e., bias ratio $>$ 1.2) for events $\leq$ 6.35 mm (0.25 in.). HRRRv3 frequency biases were closer to neutral for events $\leq$ 3.81 mm (0.15 in.), but not significantly different from those of HRRRv2 at a 95% confidence level, as determined using bootstrap resampling for ratios of event frequency [subsequent statements of confidence also use this technique (Choquet et al. 1999; Hamill 1999)].

At SNOTEL stations, we present frequency bias for events in five bins similarly defined by lower and upper bounds [1.27–6.35 mm (0.05–0.25 in.), 6.35–19.05 mm (0.25–0.75 in.), 19.05–31.75 mm (0.75–1.25 in.), 31.75–44.45 mm (1.25–1.75 in.), and 44.45–57.15 mm (1.75–2.25 in.)], represented in each graph by a central value (Fig. 5b). The lower bound is exclusive and the upper bound inclusive for all but the lowest bin, for which we use model precipitation values $\geq$ 1.27 mm (0.05 in.) and

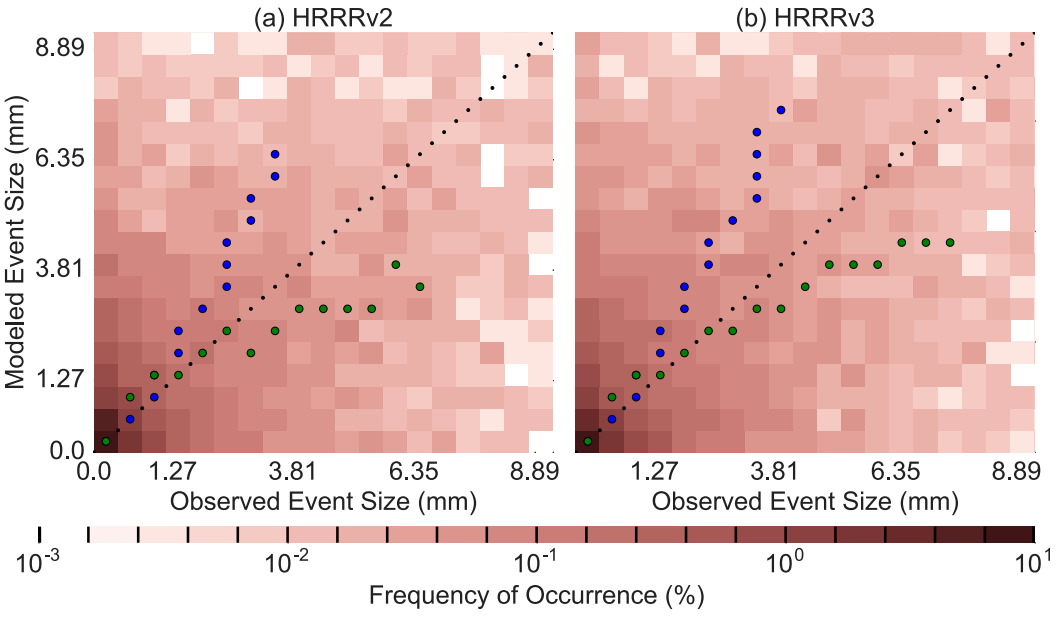

FIG. 6. Bivariate histograms of forecast and observed precipitation at ASOS/AWOS stations for (a) HRRRv2 and (b) HRRRv3. Green (blue) dots denote mean modeled (observed) event size for each observed (modeled) event size in each bin. Dots not shown for bins with <100 events.

observed precipitation values ≥ 2.54 mm (0.10 in.). Events > 57.15 mm (2.25 in.) are not presented due to the small sample size. HRRRv2 frequency biases are <1 but fall within near-neutral bounds for all events sizes except those ≤6.35 mm (0.25 in.) where underprediction occurs. HRRRv3 bias ratios are higher for all events except those ≤6.35 mm (0.25 in.), consistent with the higher mean bias ratio, with slight overprediction for events ~38.1 mm (1.5 in.), which is the only bin in which the difference is significant at a 95% confidence level.

Bivariate histograms illustrate bias if frequent event pairs fall above (overprediction) or below (underprediction) the 1:1 line and precision based on the scatter of event pairs. Ideally, most event pairs fall along or near the 1:1 line. At ASOS/AWOS stations, the

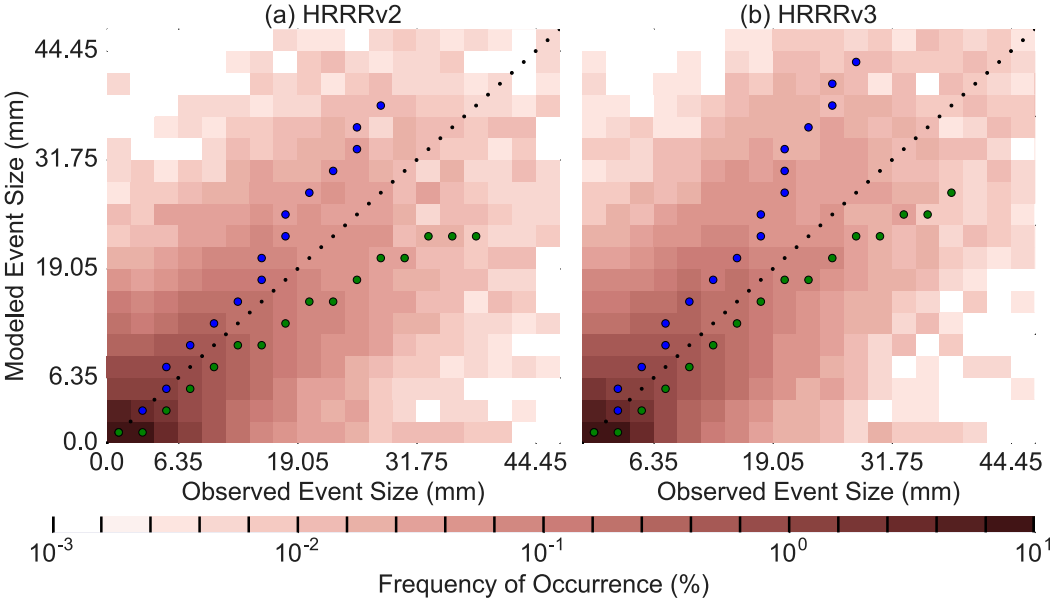

FIG. 7. As in Fig. 6, but for SNOTEL stations.

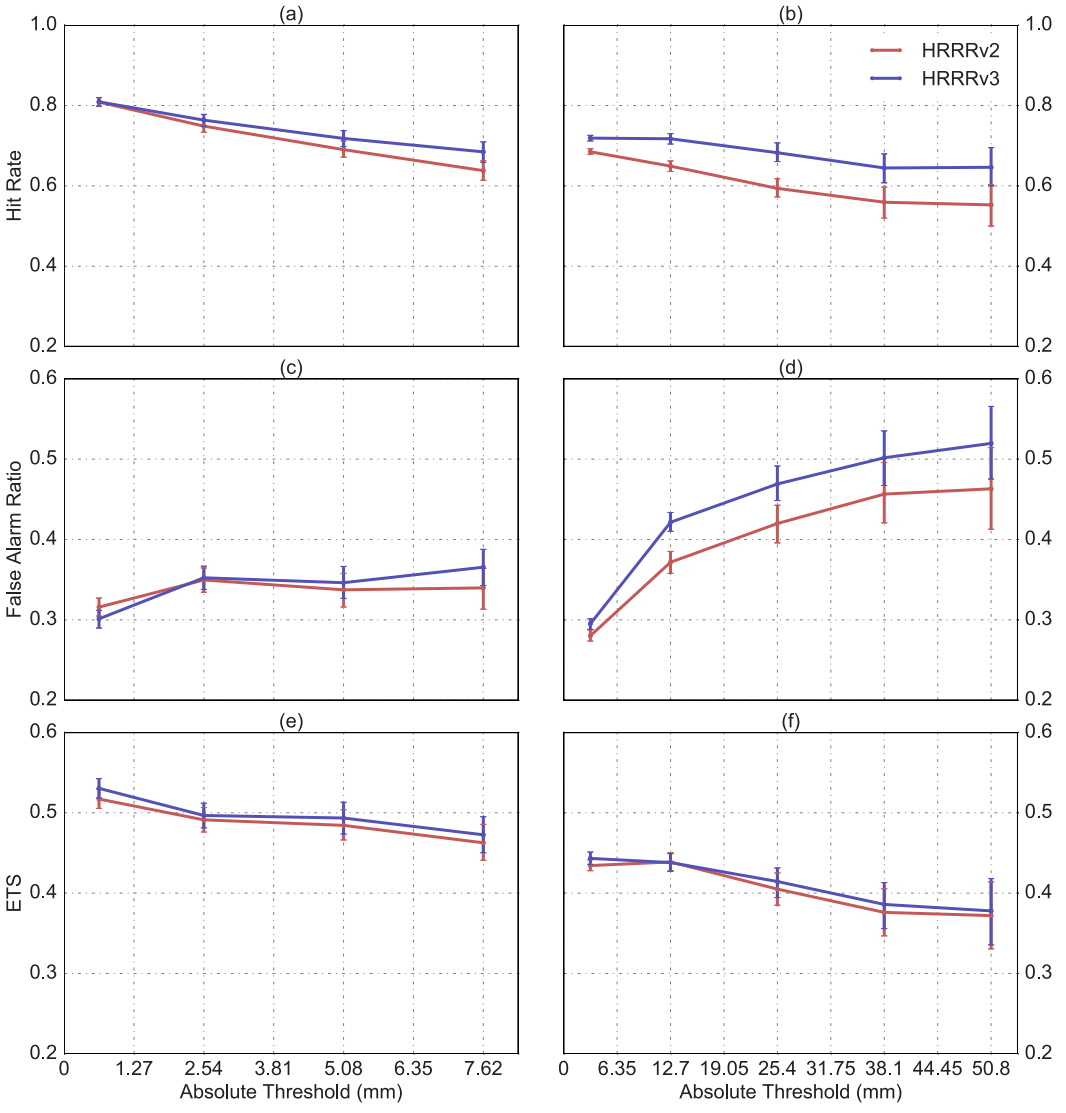

FIG. 8. HRRRv2 (red) and HRRRv3 (blue) verification metrics as functions of absolute thresholds at (left) ASOS/AWOS and (right) SNOTEL stations. (a),(b) Hit rate. (c),(d) False alarm ratio. (e),(f) Equitable threat score. Whiskers display 95% confidence intervals as determined using bootstrap resampling.

HRRRv2 bivariate histogram displays minimal skewness about the 1:1 line, which suggests near-neutral bias, but low precision, indicated by large scatter of event pairs (Fig. 6a). The HRRRv3 bivariate histogram similarly reveals minimal skewness but low precision (Fig. 6b). Thus, while the model biases were small, the large scatter indicates weak correlation between forecasts and observations, a result that may partly reflect undersampling of events at ASOS/AWOS stations. At SNOTEL stations, the HRRRv2 bivariate histogram exhibits near-neutral bias and moderate precision (Fig. 7a). The HRRRv3 bivariate histogram indicates similar performance (Fig. 7b). Altogether, the HRRRv2 and HRRRv3 bias ratios, frequency biases, and bivariate histograms indicate a near-neutral precipitation bias for total precipitation and most event sizes, with HRRRv3 slightly wetter than HRRRv2. For both, precision increases from lowland ASOS/AWOS stations to upland SNOTEL stations. Low precision at the lowland ASOS/AWOS stations may partially reflect undersampling.

We next evaluate model skill using the traditional metrics of HR, FAR, and ETS. Whereas the HR and FAR examine how well the model captures events or nonevents, the ETS measures skill relative to random forecasts (drawn from the observed climatological distribution). At ASOS/AWOS stations, as absolute threshold increases, HRRRv2 HR decreases from 0.81

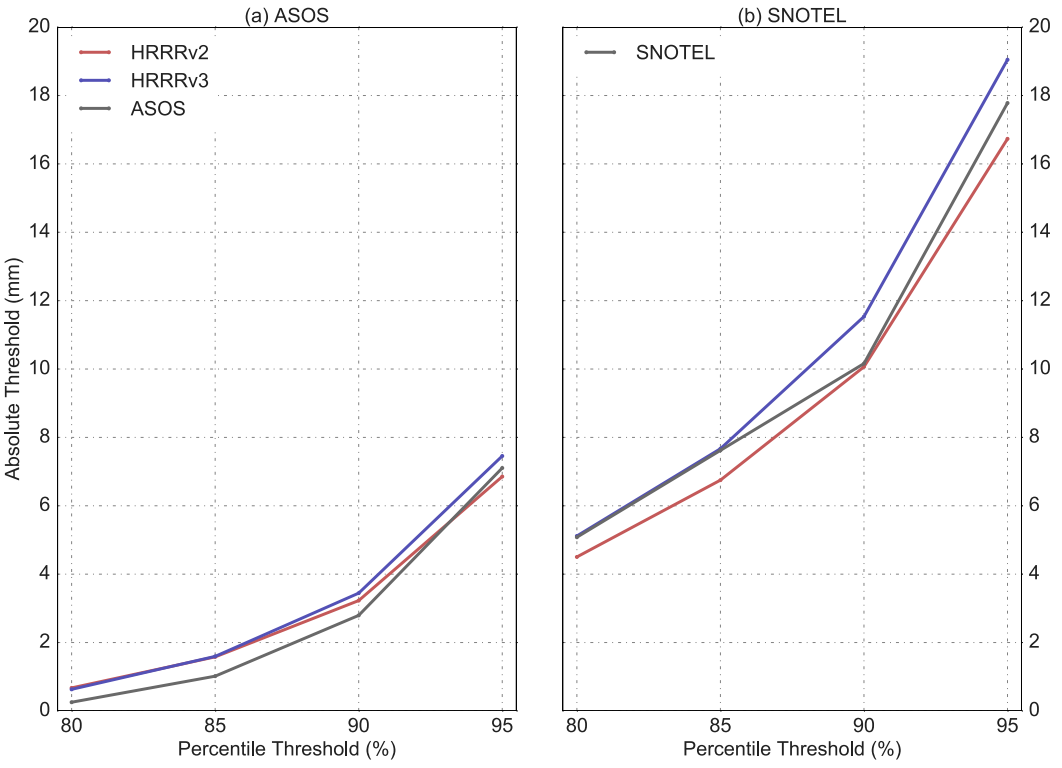

FIG. 9. Observed (gray) and forecast HRRRv2 (red) and HRRRv3 (blue) absolute and precipitation thresholds at (a) ASOS/AWOS and (b) SNOTEL stations.

to 0.64 (Fig. 8a), FAR increases from 0.32 to 0.35 (Fig. 8c), and ETS decreases from 0.52 to 0.46 (Fig. 8e). HRRRv3 HRs, FARs, and ETSs are larger in comparison at most event thresholds, although differences are not significant at a 95% confidence level. At SNOTEL stations, HRRRv2 HR decreases from 0.68 to 0.55 (Fig. 8b), FAR increases from 0.28 to 0.46 (Fig. 8d), and ETS decreases from 0.44 to 0.37 (Fig. 8f). Similar to ASOS/AWOS stations, HRRRv3 HRs and FARs are larger than those of HRRRv2 and the ETS is comparable to or slightly higher at all thresholds. Although the differences in HR and FAR are sometimes significant, specifically at lower thresholds, differences in ETS are not significant at a 95% confidence level. Comparison of ASOS/AWOS and SNOTEL results indicates higher ETS in the lowlands compared to the uplands. Although the thresholds used for each network differ, this is true for ASOS/AWOS SNOTEL thresholds that overlap.

Next, we convert absolute thresholds to percentile thresholds for each modeling system and station network according to Fig. 9. This helps to account for model bias, although such biases are small for HRRRv2 and HRRRv3. As percentile threshold increases at ASOS/AWOS stations, HRRRv2 HR decreases from 0.77 to 0.66 (Fig. 10a), FAR increases from 0.26 to 0.34 (Fig. 10c), and ETS decreases from 0.53 to 0.47 (Fig. 10e). Compared to HRRRv2, HRRRv3 HR and ETS are larger and FAR is smaller, although the differences are not significant at a 95% confidence level. As percentile threshold increases at SNOTEL stations, HRRRv2 HR decreases from 0.75 to 0.64 (Fig. 10b), FAR varies between 0.41 and 0.27 (Fig. 10d), and ETS decreases from 0.45 to 0.44 (Fig. 10f). The HRRRv3 HR and ETS are slightly higher and FAR slightly lower, although the differences are not significant at a 95% confidence level. Similar to the results for absolute thresholds, ETS is higher at the lowland ASOS/AWOS sites than the upland SNOTEL sites.

To summarize, comparison of HRRRv2 and HRRRv3 during the 2017/18 cool season indicates little change in model biases and performance characteristics. Both models were slightly wet at lowland ASOS/AWOS stations and near neutral at upland SNOTEL stations. At both ASOS/AWOS and SNOTEL stations, the HRRRv3 exhibited higher HR and ETS and lower FAR, but differences in ETS were not significant at a 95% confidence level. These results suggest a small, but statistically undiscernible improvement from HRRRv2

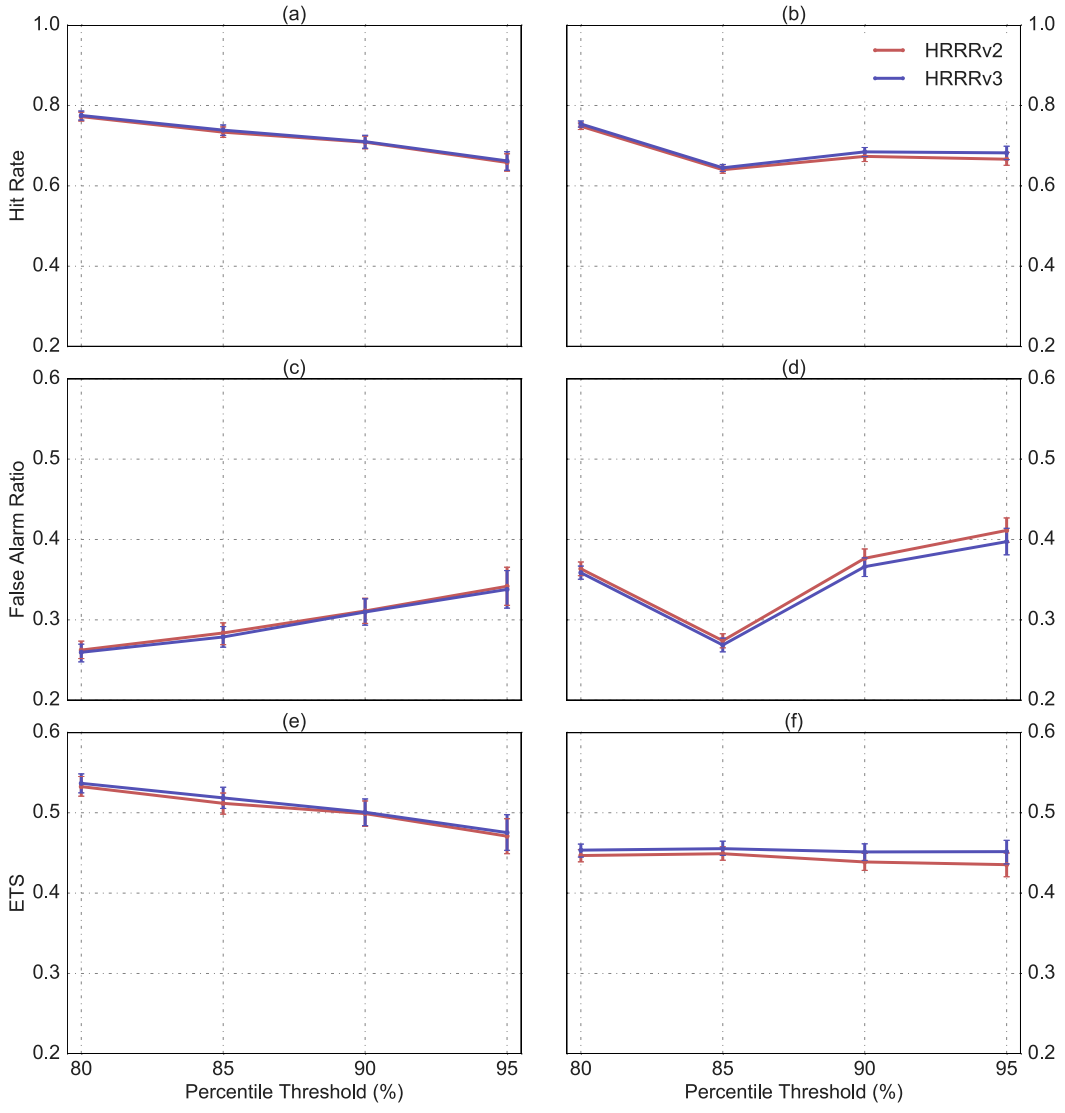

FIG. 10. As in Fig. 8, but for precipitation thresholds.

to HRRRv3. We hypothesize that these differences are likely not distinguishable to operational forecasters. ETS was higher at ASOS/AWOS sites than upland SNOTEL sites for both absolute and percentile thresholds.

c. *GFSv14, GFSv15.0, and HRES*

At ASOS/AWOS stations, GFSv14 bias ratios indicate that forecasts tended to be wet, with a mean bias ratio of 1.65 on day 1 that decreases slightly to 1.57 on day 3 (Figs. 11a,b). There are large standard deviations on day 1 (1.05) and day 3 (1.02), which reflect large wet biases at many stations. GFSv15.0 mean bias ratios are slightly higher at 1.77 on day 1 and 1.65 on day 3 (Figs. 11c,d), with comparable standard deviations. HRES forecasts were the wettest, with mean day-1 and day-3 bias ratios of 1.80 and 1.91, respectively, and comparable standard deviations (Figs. 11e,f). In contrast, at SNOTEL stations, mean GFSv14 day-1 and day-3 bias ratios are 0.99 and 0.97, respectively, with substantially lower standard deviations (Figs. 12a,b). GFSv15.0 forecasts were similar, with day-1 and day-3 bias ratios of 1.00 and 0.96, respectively (Figs. 12c,d). HRES forecasts exhibited a weak dry bias, with mean day-1 and day-3 bias ratios of 0.88 and 0.91, respectively (Figs. 12e,f).

Consistent with the high bias ratios, all three models overpredicted the frequency of day-1 and day-3 precipitation events at ASOS/AWOS stations for all event sizes (Fig. 13a). This problem was most acute in

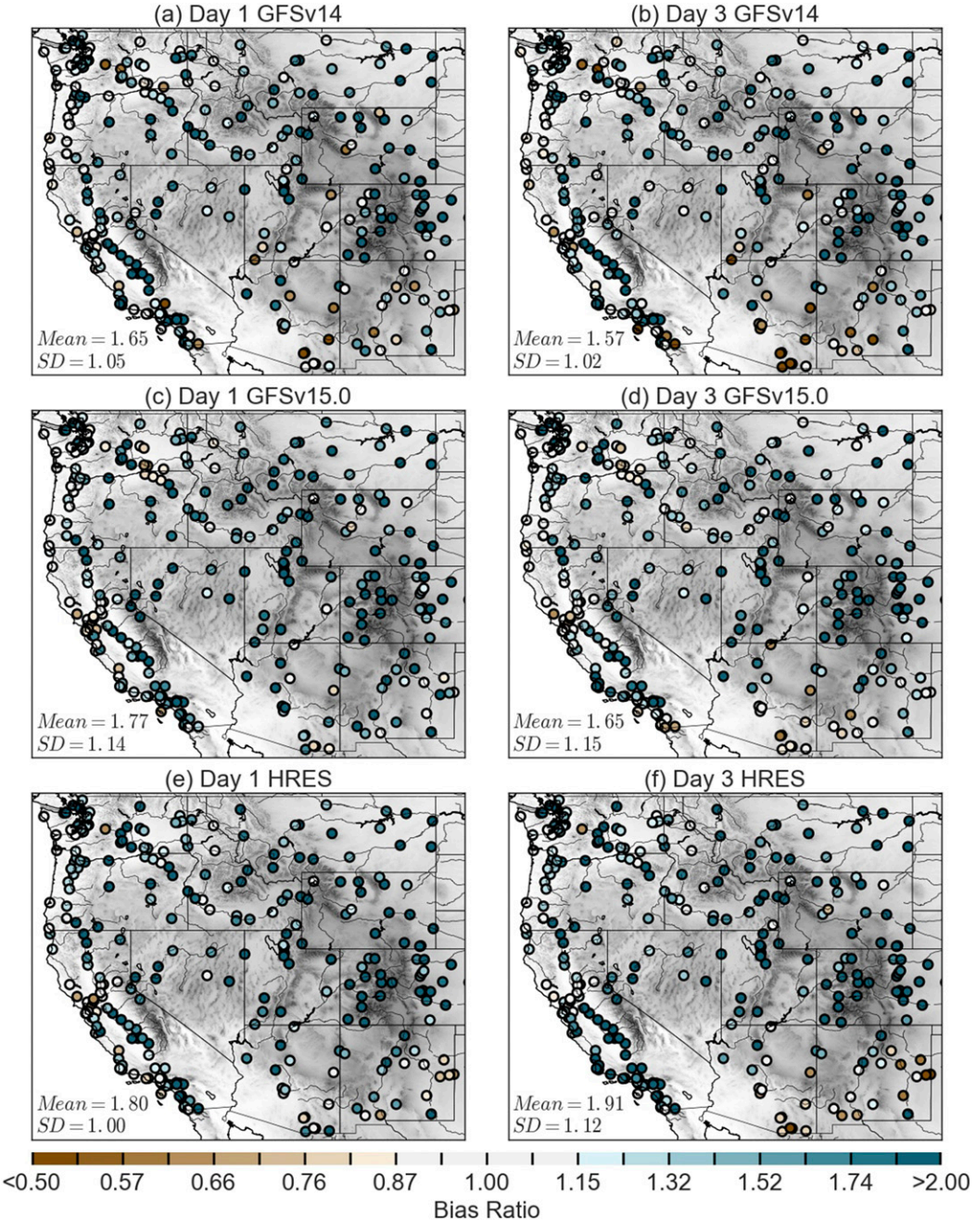

FIG. 11. (a) Day-1 GFSv14, (b) day-3 GFSv14, (c) day-1 GFSv15.0, (d) day-3 GFSv15.0, (e) day-1 HRES, and (f) day-3 HRES bias ratios at ASOS/AWOS stations with 30 arc-s topography (as in Fig. 1). Mean and standard deviation (SD) annotated.

HRES forecasts, consistent with the larger HRES wet bias. At SNOTEL stations, all three models exhibited near-neutral or marginally low-frequency biases on day 1 and day 3 for all event sizes (Fig. 13b). Underprediction of event frequency was more apparent at higher thresholds and increased from the GFSv15.0 to GFSv14 to HRES.

Bivariate histograms illustrate that GFSv14 event pairs at ASOS/AWOS stations were skewed above the 1:1 line, which is consistent with the aforementioned wet bias (Figs. 14a,b). Furthermore, the large scatter of event pairs reflects low precision. The GFSv15.0 and HRES displayed similar skewness and scatter at ASOS/AWOS stations (Figs. 14c–f). At SNOTEL

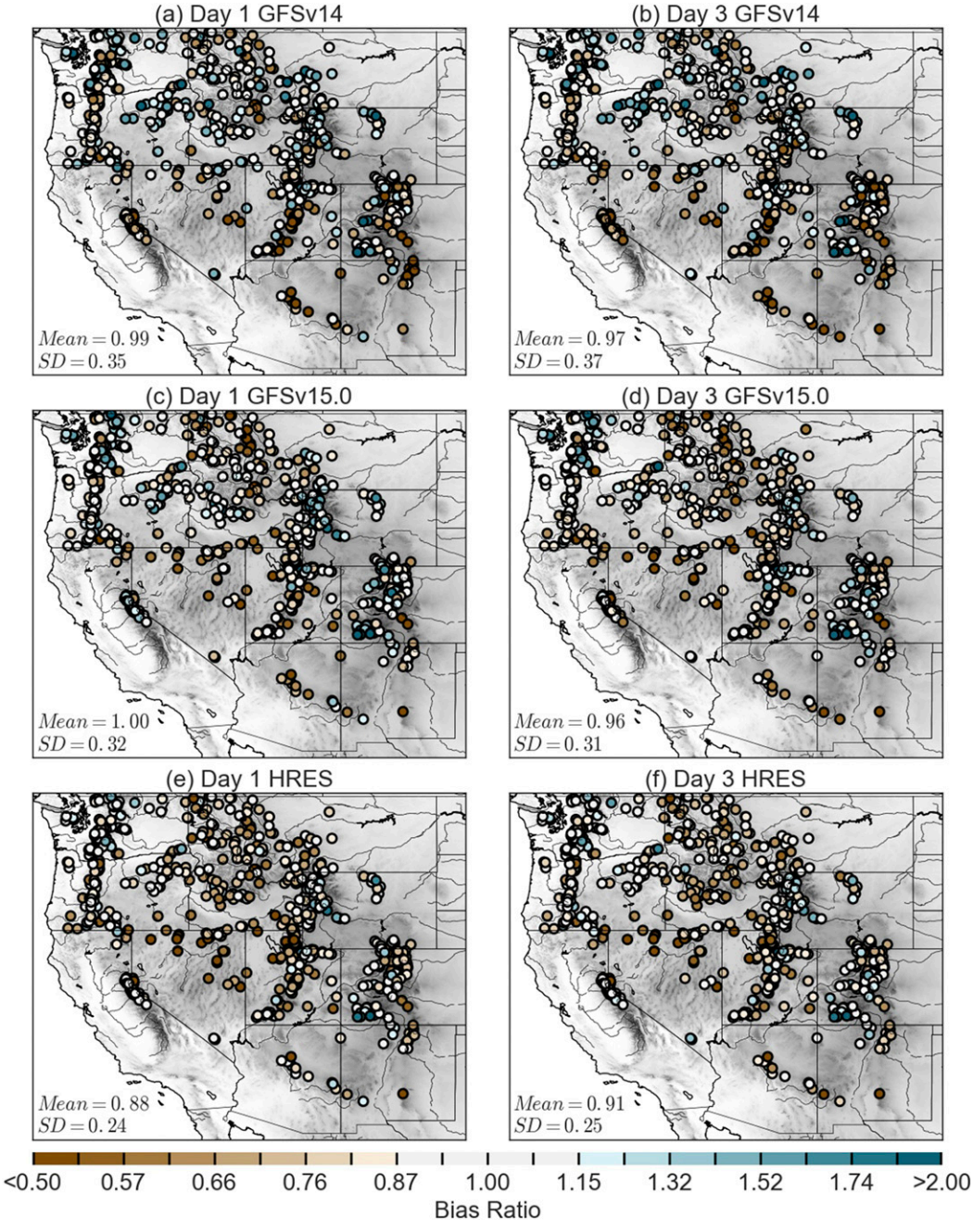

FIG. 12. As in Fig. 11, but for SNOTEL stations.

stations, the GFSv14 bivariate histogram exhibited minimal skewness and, for small events, small scatter, indicating near-neutral bias and moderately high precision (Fig. 15a). Precision declined, however, for larger events and for longer lead times (cf. Figs. 15a,b). The GFSv15.0 bivariate histograms exhibit similar characteristics (Figs. 15c,d). HRES, however, skewed below the 1:1 line and thus displayed slight underprediction, consistent with its weak dry bias (Figs. 15e,f). Overall, these results indicate that all three global models produce excessive lowland precipitation, but the bias is neutral or dry in upland regions, with the HRES featuring the largest upland underprediction, especially for larger events.

HR and ETS are generally highest for HRES and lowest for GFSv14 at both ASOS/AWOS and SNOTEL stations on day 1 and day 3 (Figs. 16a,b,e,f). For FAR, differences between the models are modest at ASOS/AWOS

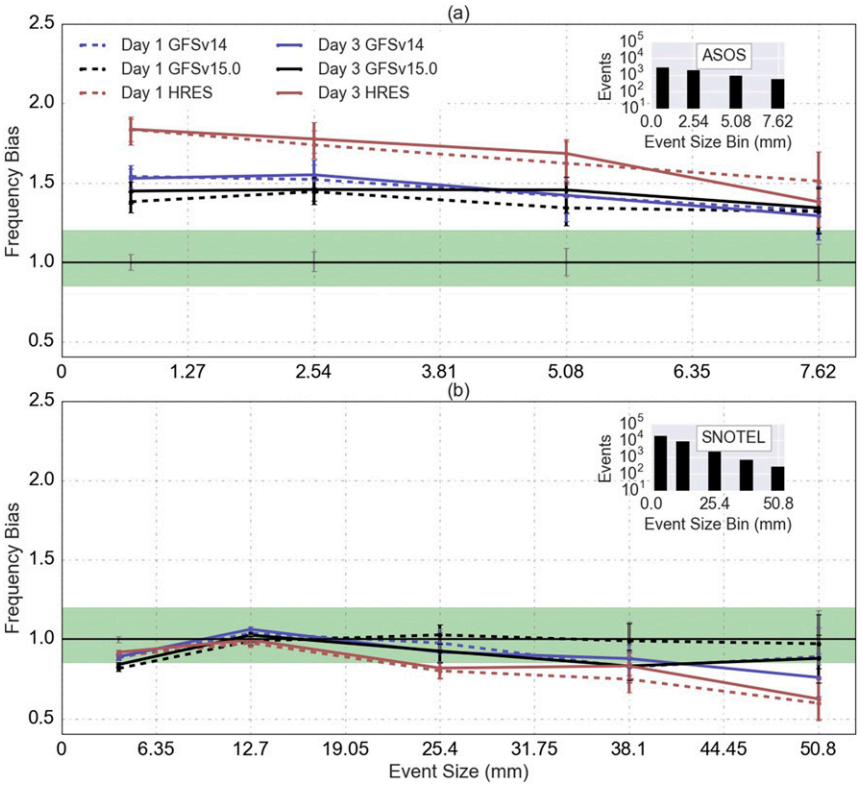

FIG. 13. Day-1 (dashed) and day-3 (solid) GFSv14 (blue), GFSv15.0 (black), and HRES (red) frequency bias as a function of event size at (a) ASOS/AWOS and (b) SNOTEL stations. Number of events sampled into each bin shown in inset histograms. Green band shows 0.85–1.20 range defined as near neutral by the authors. Whiskers display 95% confidence intervals as determined using bootstrap resampling.

stations, but the drier HRES leads to much lower values at SNOTEL stations, especially on day 1 (Figs. 16c,d). Focusing on ETS as an overall indicator of model performance, on day 1, the HRES produces the highest ETS for all but the smallest [≤1.27 mm (0.05 in.)] events at ASOS/AWOS stations and all events at SNOTEL stations, with the improvement relative to GFSv14 and GFSv15.0 significant at a 95% confidence level in several size bins (Figs. 16e,d). Although ETS declines by day 3, the gap between HRES and GFSv15.0 is smaller at both ASOS/AWOS and SNOTEL stations and not significant at a 95% confidence level for all event sizes. The gap between GFSv15.0 and GFSv14 also increases from day 1 to day 3 for most event sizes. Similar to the results for the HRRR versions, ETS for all three global models is higher at lowland ASOS/AWOS sites than upland SNOTEL sites, including thresholds that overlap.

Figure 17 illustrates the relationship between absolute thresholds and percentile thresholds for the three global models. Validating based on percentile thresholds helps account for model bias, which is more significant for the three global models than the HRRR. Based on these percentile thresholds, the HRES produces the highest HR, lowest FAR, and highest ETS on day 1 and day 3 for all event sizes at both ASOS/AWOS and SNOTEL stations (Fig. 18). The difference between GFSv15.0 and GFSv14 is small on day 1, especially at ASOS/AWOS stations, but increases by day 3, with the former producing a higher HR, lower FAR, and higher ETS in all categories. For ETS, the difference between HRES and GFSv15.0 or GFSv14 is statistically significant in nearly all thresholds on day 1 at ASOS/AWOS stations and all thresholds at SNOTEL stations. Consistent with the ETS for absolute thresholds, however, GFSv15.0 closes the gap by day 3. The gap between GFSv15.0 and GFSv14 also increases from day 1 to day 3, for which it is significant at a 95% confidence level for all event sizes at SNOTEL stations. Similar to the results for absolute thresholds, ETS is higher at lowland ASOS/AWOS sites than upland SNOTEL sites.

In summary, all three global models produce too much and too frequent precipitation at lowland

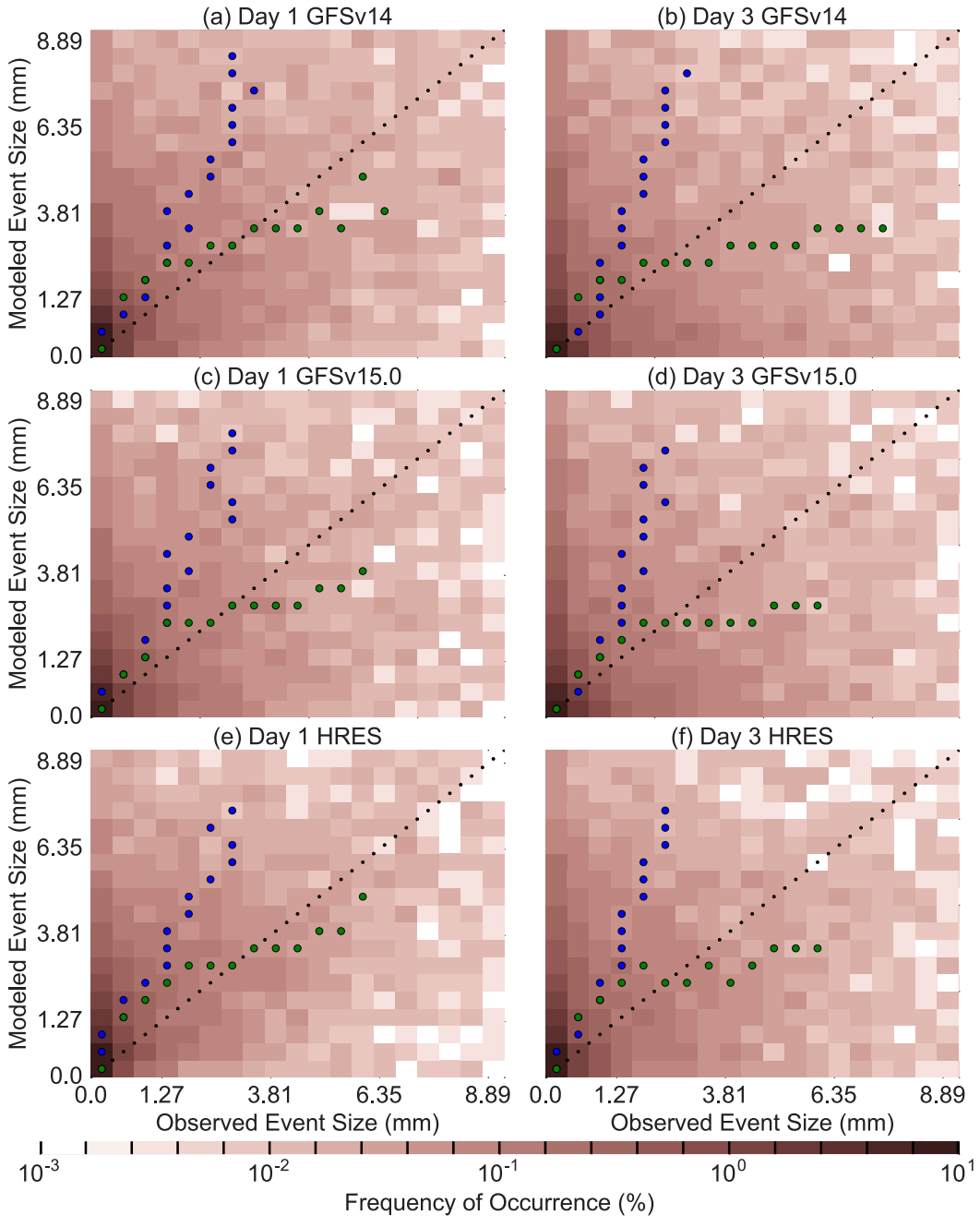

FIG. 14. Bivariate histograms of forecast and observed precipitation at ASOS/AWOS stations for (a) day-1 GFSv14, (b) day-3 GFSv14, (c) day-1 GFSv15.0, (d) day-3 GFSv15.0, (e) day-1 HRES, and (f) day-3 HRES. Green (blue) dots denote mean modeled (observed) event size for each observed (modeled) event size in each bin. Dots not shown for bins with <100 events.

ASOS/AWOS stations. Biases at upland SNOTEL stations are closer to neutral or dry, with the HRES tending to produce too little precipitation overall and too infrequent larger events. Model skill scores illustrate superior performance of the HRES at both lowland ASOS/AWOS stations and upland SNOTEL stations, especially if one validates based on percentiles, which helps account for the HRES dry bias. The difference between GFSv15.0 and GFSv14 is small on day 1 but increases by day 3 when the former has also closed the gap relative to HRES. Based on the traditional metrics used here, the shorter-range (day 1 and day 3)

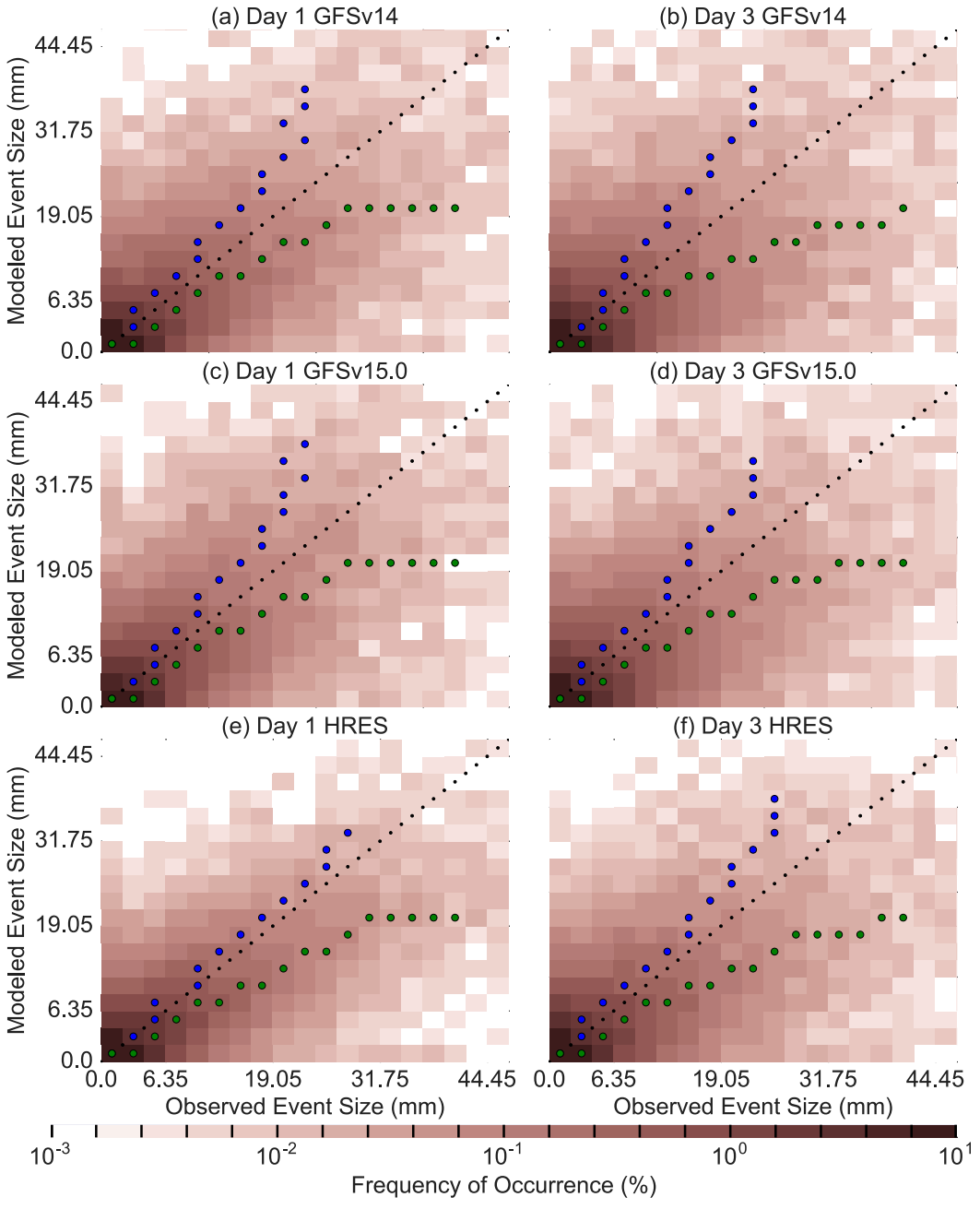

FIG. 15. As in Fig. 14, but for SNOTEL stations.

precipitation forecasts produced by GFSv15.0 produce comparable to superior forecasts to GFSv14, although they lag HRES. ETS for all three global models is higher at lowland ASOS/AWOS sites than upland SNOTEL sites for both absolute and percentile thresholds.

4. Conclusions

This study has examined the performance of newly upgraded NCEP operational models compared to their predecessors focusing on precipitation over the western CONUS during the 2017/18 cool season. Results of the evaluation can be condensed into two principal conclusions. First, changes in bias and performance between HRRRv2 and HRRRv3 are small. In the case of performance, HRRRv3 produced marginally higher ETS at lowland and upland stations, although the difference was not significant at a 95% confidence level. Second, as evaluated using traditional metrics, GFSv15.0 produces forecasts that are

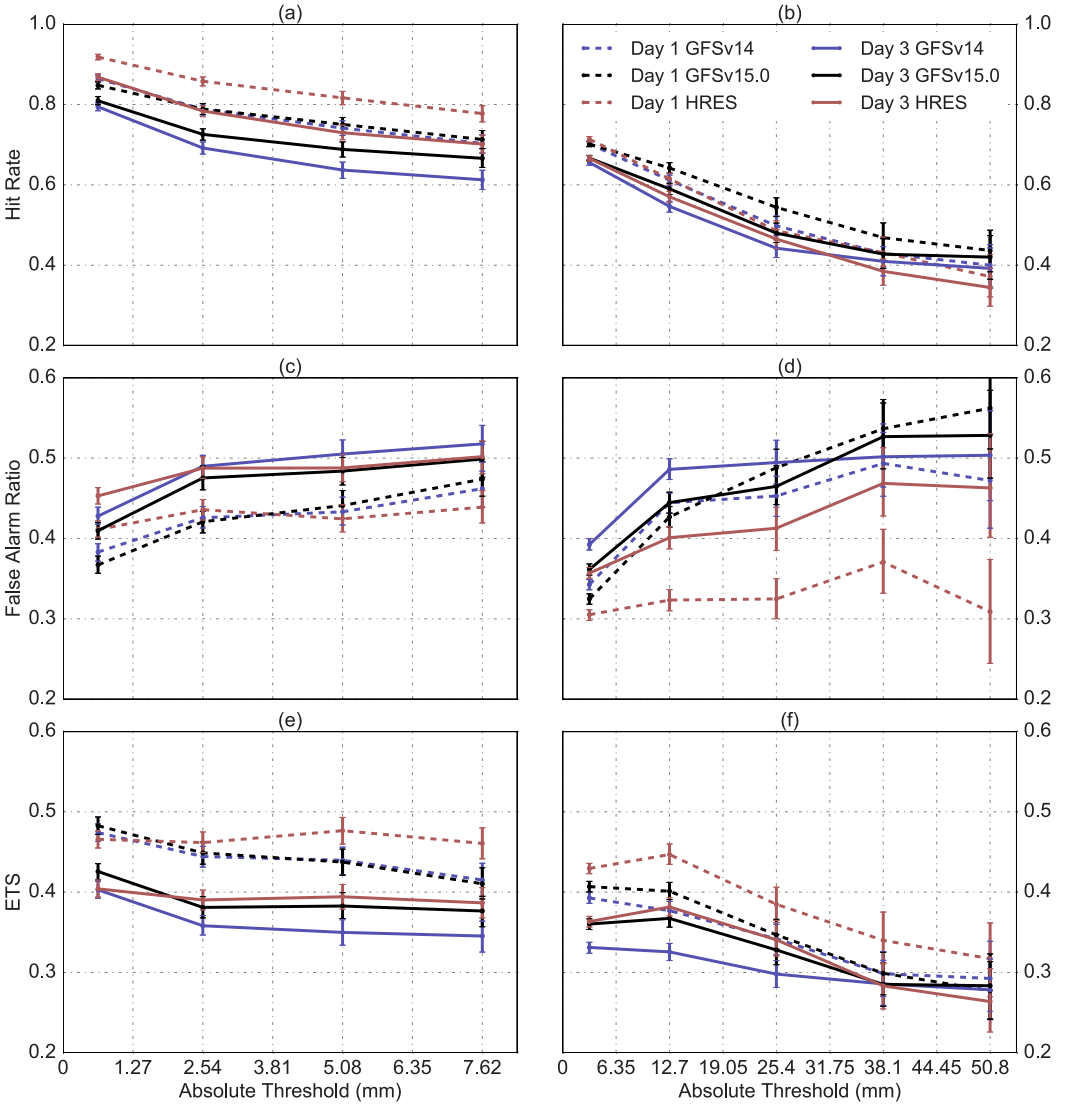

FIG. 16. Day-1 (dashed) and day-3 (solid) GFSv14 (blue), GFSv15.0 (black), and HRES (red) verification metrics as functions of absolute thresholds at (left) ASOS/AWOS and (right) SNOTEL stations. (a),(b) Hit rate. (c),(d) False alarm ratio. (e),(f) Equitable threat score. Whiskers display 95% confidence intervals as determined using bootstrap resampling.

comparable to (day 1) or superior to (day 3) GFSv14, but that still lag HRES, although the gap closes from day 1 to day 3. All three global models (GFSv15.0, GFSv14, and HRES) produce too much and too frequent lowland precipitation, but exhibit near-neutral or dry biases in upland regions, with the HRES producing the largest underprediction of larger upland precipitation events. These elevation-dependent biases may reflect insufficient terrain representation, which yields weakened orographic influences on precipitation. Superior performance of the HRES is especially apparent if one verifies using event percentiles, which helps account for these biases. Operational forecasters should be aware of the general biases described here, but also that there are variations by location and event size.

Comparison of results at ASOS/AWOS stations with SNOTEL stations indicates that ETS is higher in the lowlands than the uplands for both versions of the HRRR and all versions of the global models. This suggests a decrease in skill from the lowlands to the uplands across several forecast systems. We note, however, that bias ratios and frequency biases are also higher over the lowlands than the uplands,

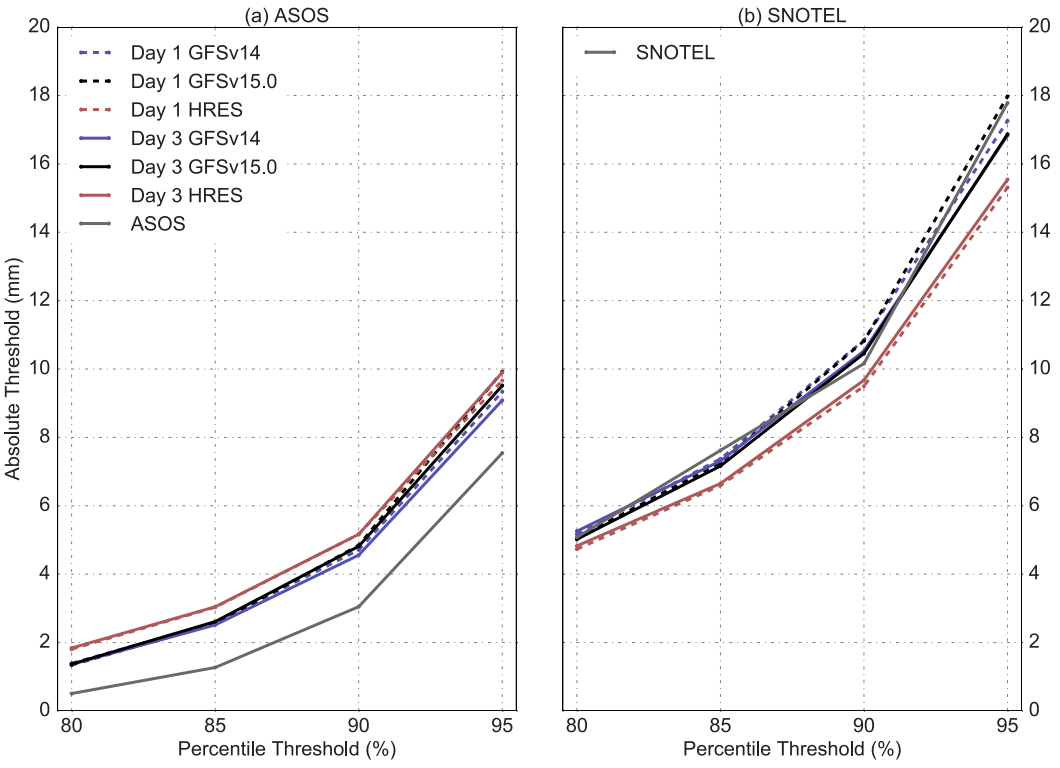

FIG. 17. Observed (gray) and forecast day-1 (dashed) and day-3 (solid) GFSv14 (blue), GFSv15.0 (black), and HRES (red) absolute and percentile precipitation thresholds at (a) ASOS/AWOS and (b) SNOTEL stations.

and wetter forecasts tend to produce higher ETS (e.g., Mason 1989; Hamill 1999). In addition, metrics like ETS are dependent on climatology (Hamill and Juras 2006), and the precipitation climatologies of the lowlands and uplands are fundamentally different. Further work is needed to better understand the causes of lower ETS in the mountains and whether or not this reflects inherent differences in predictability between lowland and mountain cool-season precipitation events.

These results are, however, based on a single cool season characterized by near- or slightly above-average precipitation in the northwest CONUS and below-average precipitation in the southwest CONUS. Thus, precipitation events in the northwest CONUS have a strong influence on overall results. Nevertheless, the HRRR and GFS biases described here for upland SNOTEL stations are broadly consistent with those identified by Gowan et al. (2018) in the then operational versions of the HRRR and GFS during the 2016/17 cool season (Gowan et al. 2018 did not examine HRRR and GFS performance at lowland ASOS/AWOS stations). Large station-by-station variations in bias ratio were identified at ASOS/AWOS stations, but likely reflect undersampling. Although a multi-cool-season model comparison study is desirable, it is not always possible with operational modeling systems. GFSv15.0 reforecasts are, however, available for three cool seasons, although for brevity we focused this paper on the 2017/18 cool season given that HRRRv2 and HRRRv3 were only available that cool season.

This study also utilized observations from the ASOS/AWOS and SNOTEL networks, which enables comparison of model performance in lowland and upland areas. Both station types, however, likely experience undercatch, which is not accounted for here, and the quality control and assessment of 24-h precipitation amounts at SNOTEL stations is difficult and lacks data precision. A major advantage of the SNOTEL network, however, is its high density in mountain areas that are poorly sampled by radar and exhibit large uncertainties in gridded precipitation analyses. Future validation studies over the western CONUS should continue to leverage the SNOTEL network (and potentially other mountain observing stations) to better identify model biases and performance characteristics in upland areas where forecasts are critical for recognizing impacts related to

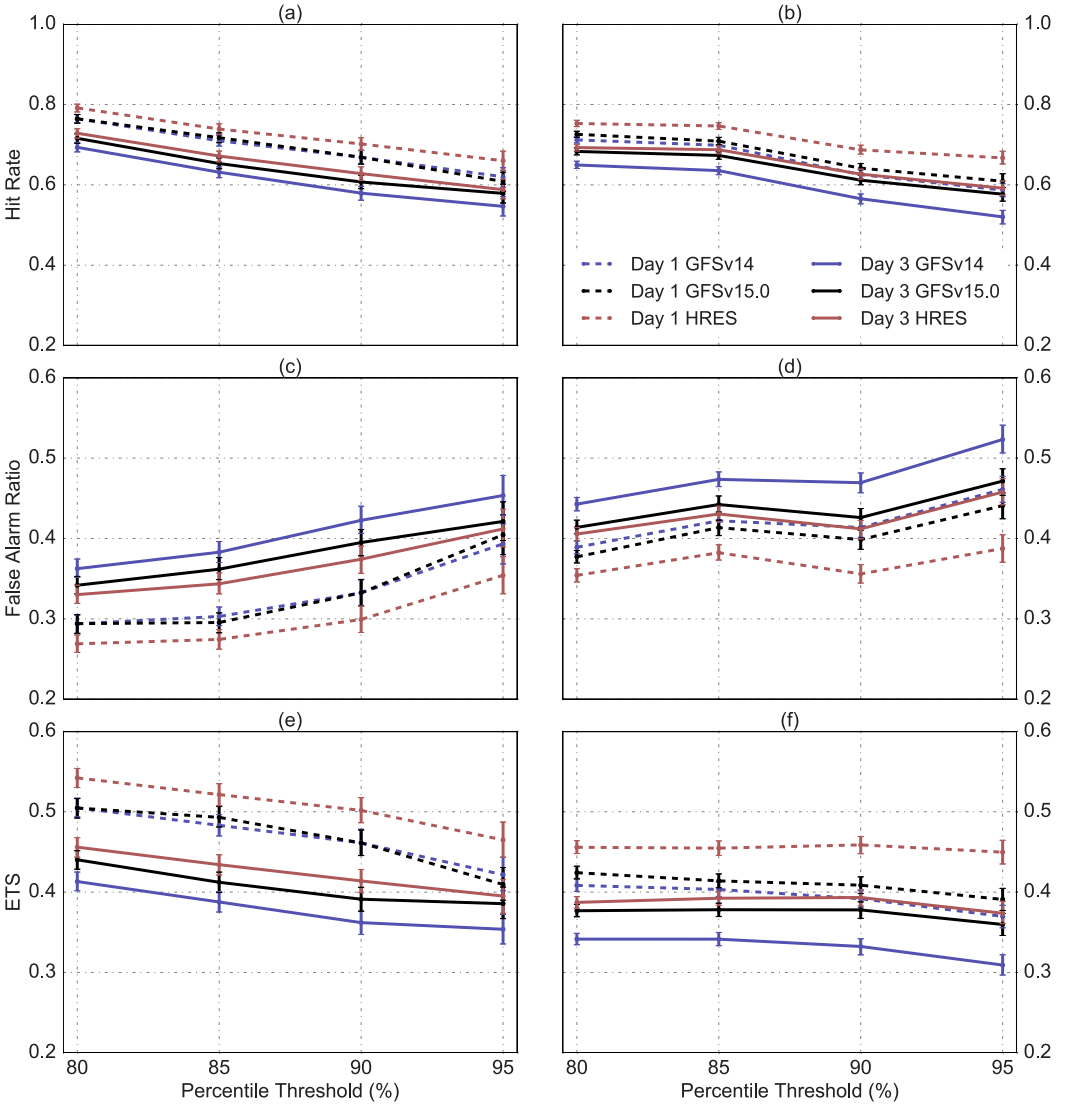

FIG. 18. As in Fig. 16, but for percentile thresholds.

flooding, debris flows, avalanches, and road maintenance and safety.

Acknowledgments. We thank Trevor Alcott, Thomas Haiden, the Global Systems Division at the NOAA/ESRL, and the NOAA/NCEP/EMC for their assistance in providing model data. We also thank the NRCS for access to SNOTEL data, Synoptic Data for access to ASOS/AWOS data, the PRISM climate group at Oregon State University for access to gridded climatological precipitation analyses, and the University of Utah Center for High Performance Computing for providing computational resources and support. Tom Gowan and Peter Veals provided suggestions and programming assistance and Court Strong, John Horel, Trevor Alcott, and three anonymous reviewers provided comments and suggestions that improved the manuscript. This article is based on research supported by the NOAA/National Weather Service CSTAR Program through Grant NA17NWS4680001. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect those of the NOAA/National Weather Service.

REFERENCES

Andrey, J., B. Mills, and J. Vandermolen, 2001: Weather information and road safety. Institute for Catastrophic Loss Reduction Paper Series 15, ICLR, 36 pp., http://0361572.netsolhost.com/images/Weather_information_and_road_safety.pdf.

Barbero, R., J. T. Abatzoglou, and H. J. Fowlerm, 2019: Contribution of large-scale midlatitude disturbances to hourly precipitation extremes in the United States. *Climate Dyn.*, **52**, 197–208, https://doi.org/10.1007/s00382-018-4123-5.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Birkeland, K. W., and C. J. Mock, 2001: The major snow avalanche cycle of February 1986 in the western United States. *Nat. Hazards*, **24**, 75–95, https://doi.org/10.1023/A:1011192619039.

Chen, J.-H., and S.-J. Lin, 2013: Seasonal predictions of tropical cyclones using a 25-km-resolution general circulation model. *J. Climate*, **26**, 380–398, https://doi.org/10.1175/JCLI-D-12-00061.1.

——, X. Chen, S.-J. Lin, L. Magnusson, M. Bender, L. Zhou, and S. Rees, 2018: Tropical cyclones in the global 8.5km GFDL fvGFS. *33rd Conf. on Hurricanes and Tropical Meteorology*, Ponte Vedra, FL, Amer. Meteor. Soc., 9B.4, https://ams.confex.com/ams/33HURRICANE/webprogram/Paper339827.html.

Choquet, D., P. L'Ecuyer, and C. Léger, 1999: Bootstrap confidence intervals for ratios of expectations. *ACM Trans. Model. Comput. Simul.*, **9**, 326–348, https://doi.org/10.1145/352222.352224.

Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016: Convection-permitting models: A step-change in rainfall forecasting. *Meteor. Appl.*, **23**, 165–181, https://doi.org/10.1002/met.1538.

Collins, W. D., and Coauthors, 2004: Description of the NCAR Community Atmosphere Model (CAM 3.0). NCAR Tech. Note NCAR/TN-464+STR, 214 pp., https://doi.org/10.5065/D63N21CH.

Craven, P. C., and Coauthors, 2018: Overview of National Blend of Models version 3.1. Part I: Capabilities and an outlook for future upgrades. *25th Conf. on Probability and Statistics*, Austin, TX, Amer. Meteor. Soc., 7.3, https://ams.confex.com/ams/98Annual/webprogram/Paper325347.html.

Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.*, **33**, 140–158, https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

——, W. P. Gibson, G. H. Taylor, M. K. Doggett, and J. I. Smith, 2007: Observer bias in daily precipitation measurements at United States cooperative network stations. *Bull. Amer. Meteor. Soc.*, **88**, 899–912, https://doi.org/10.1175/BAMS-88-6-899.

Dey, S. R. A., G. Leoncini, N. M. Roberts, R. S. Plant, and S. Migliorini, 2014: A spatial view of ensemble spread in convection-permitting ensembles. *Mon. Wea. Rev.*, **142**, 4091–4107, https://doi.org/10.1175/MWR-D-14-00172.1.

Durre, I., M. F. Squires, R. S. Vose, X. Yin, A. Arguez, and S. Applequist, 2013: NOAA's 1981–2010 U.S. climate normals: Monthly precipitation, snowfall, and snow depth. *J. Appl. Meteor. Climatol.*, **52**, 2377–2395, https://doi.org/10.1175/JAMC-D-13-051.1.

ECMWF, 2019: Changes in ECMWF model. ECMWF, accessed 23 October 2019, https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model.

Federal Aviation Administration, 2017: Automated Weather Observing Systems (AWOS) for non-federal applications. Advisory Circular 150/5220-16E, FAA, 77 pp., https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_150_5220-16E.pdf.

——, 2020: Surface weather observing sations. FAA, accessed 18 February 2020, https://www.faa.gov/air_traffic/weather/asos/.

Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, https://doi.org/10.1175/WAF-D-17-0144.1.

Greeney, C. M., M. D. Gifford, and M. L. Salyards, 2005: Winter test of production all-weather precipitation accumulation gauge for ASOS 2003–2004. *Ninth Symp. on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IOAS-AOLS)*, San Diego, CA, Amer. Meteor. Soc., 8.3, https://ams.confex.com/ams/Annual2005/techprogram/paper_82895.htm.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

Hashimoto, H., W. Wang, F. S. Melton, A. L. Moreno, S. Ganguly, A. R. Michaelis, and R. R. Nemani, 2019: High-resolution mapping of daily climate variables by aggregating multiple spatial data sets with the random forest algorithm over the conterminous United States. *Int. J. Climatol.*, **39**, 2964–2983, https://doi.org/10.1002/joc.5995.

Hazelton, A. T., L. M. Harris, and S.-J. Lin, 2018: Evaluation of tropical cyclone structure forecasts in a high-resolution version of the multiscale GFDL fvGFS model. *Wea. Forecasting*, **33**, 419–442, https://doi.org/10.1175/WAF-D-17-0140.1.

Horel, J., and Coauthors, 2002: Mesowest: Cooperative mesonets in the western United States. *Bull. Amer. Meteor. Soc.*, **83**, 211–225, https://doi.org/10.1175/1520-0477(2002)083<0211:MCMITW>2.3.CO;2.

Ikeda, K., and Coauthors, 2010: Simulation of seasonal snowfall over Colorado. *Atmos. Res.*, **97**, 462–477, https://doi.org/10.1016/j.atmosres.2010.04.010.

James, C. N., and R. A. Houze Jr.., 2005: Modification of precipitation by coastal orography in storms crossing northern California. *Mon. Wea. Rev.*, **133**, 3110–3131, https://doi.org/10.1175/MWR3019.1.

Lewis, W. R., W. J. Steenburgh, T. I. Alcott, and J. J. Rutz, 2017: GEFS precipitation forecasts and the implications of statistical downscaling over the western United States. *Wea. Forecasting*, **32**, 1007–1028, https://doi.org/10.1175/WAF-D-16-0179.1.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

Martinaitis, S. M., S. B. Cocks, Y. Qi, and B. T. Kaney, 2015: Understanding winter precipitation impacts on automated gauge observations within a real-time system. *J. Hydrometeor.*, **16**, 2345–2363, https://doi.org/10.1175/JHM-D-15-0020.1.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteor. Mag.*, **37**, 75–81.

——, 2003: Binary events. *Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley and Sons, 37–76.

McClung, T., 2014: Technical implementation notice 14-46, amended. TIN14-46, National Weather Service, 8 pp., https://www.weather.gov/media/notification/tins/tin14-46gfs_aab.pdf.

Minder, J. R., D. R. Durran, G. H. Roe, and A. M. Anders, 2008: The climatology of small-scale orographic precipitation over the Olympic Mountains: Patterns and processes. *Quart. J. Roy. Meteor. Soc.*, **134**, 817–839, https://doi.org/10.1002/qj.258.

Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–354, https://doi.org/10.1175/2009WAF2222260.1.

Myrick, D., 2017: Global Forecast Systems (GFS) upgrade: Effective July 19, 2017. NWS Service Change Notice 17-67, accessed 29 March 2019, https://vlab.ncep.noaa.gov/web/gfs/past-implementations.

——, 2018: Upgrade to the RAP and HRRR analysis and forecast system. NWS Service Change Notice 18-58, 6 pp., https://www.weather.gov/media/notification/pdfs/scn18-58rap_hrrr.pdf.

Nakanishi, M., and H. Niino, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteor. Soc. Japan*, **87**, 895–912, https://doi.org/10.2151/jmsj.87.895.

NOAA, 2018: The High-Resolution Rapid Refresh (HRRR). Accessed 29 March 2019, https://rapidrefresh.noaa.gov/hrrr/.

NWS, 2009: AWPAG implementation sites. Accessed 29 March 2019, https://www.weather.gov/media/asos/ASOS%20Implementation/AWPAG.xls.

——, 2016: The Global Forecast System (GFS)—Global Spectral Model (GSM). Accessed 29 March 2019, https://www.emc.ncep.noaa.gov/GFS/doc.php.

Parker, L. E., and J. T. Abatzoglou, 2016: Spatial coherence of extreme precipitation events in the northwestern United States. *Int. J. Climatol.*, **36**, 2451–2460, https://doi.org/10.1002/joc.4504.

Prein, A. F., and Coauthors, 2015: A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Rev. Geophys.*, **53**, 323–361, https://doi.org/10.1002/2014RG000475.

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

Roberts, C. D., R. Senan, F. Molteni, S. Boussetta, M. Mayer, and S. P. E. Keeley, 2018: Climate model configurations of the ECMWF Integrated Forecasting System (ECMWF-IFS cycle 43r1) for HighResMIP. *Geosci. Model Dev.*, **11**, 3681–3712, https://doi.org/10.5194/gmd-11-3681-2018.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Rutz, J. J., W. J. Steenburgh, and F. M. Ralph, 2014: Climatological characteristics of atmospheric rivers and their inland penetration over the western United States. *Mon. Wea. Rev.*, **142**, 905–921, https://doi.org/10.1175/MWR-D-13-00168.1.

Seeherman, J., and Y. Liu, 2015: Effects of extraordinary snowfall on traffic safety. *Accid. Anal. Prev.*, **81**, 194–203, https://doi.org/10.1016/j.aap.2015.04.029.

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

Simmons, A. J., and R. Strüfing, 1983: Numerical forecasts of stratospheric warming events using a model with hybrid vertical coordinate. *Quart. J. Roy. Meteor. Soc.*, **109**, 81–111, https://doi.org/10.1002/qj.49710945905.

Steenburgh, W. J., 2003: One hundred inches in one hundred hours: Evolution of a Wasatch mountain winter storm cycle. *Wea. Forecasting*, **18**, 1018–1036, https://doi.org/10.1175/1520-0434(2003)018<1018:OHIIOH>2.0.CO;2.

Tallapragada, V., and F. Yang, 2018: Next Global Forecast System (GFS). WMO Commission for Basic Systems, Meeting of the CBS (DPFS) Expert Team on Operational Weather and Forecasting Process and Support (OWFPS) Summary, accessed 29 August 2019, https://www.wmo.int/pages/prog/www/BAS/CBS-meetings.html.

Touma, D., A. M. Michalak, D. L. Swain, and N. S. Diffenbaugh, 2018: Characterizing the spatial scales of extreme daily precipitation in the United States. *J. Climate*, **31**, 8023–8037, https://doi.org/10.1175/JCLI-D-18-0019.1.

USDA, 2014: Data management. Part 622 Snow Survey and Water Supply Forecasting National Engineering Handbook, USDA, 30 pp., https://directives.sc.egov.usda.gov/OpenNonWebContent.aspx?content=35529.wba.

Weller, H., H. G. Weller, and A. Fournier, 2009: Voronoi, Delaunay, and block-structured mesh refinement for solution of shallow-water equations on the sphere. *Mon. Wea. Rev.*, **137**, 4208–4224, https://doi.org/10.1175/2009MWR2917.1.

Yang, F., 2018: GDAS/GFS v15.0.0 upgrades for Q2FY2019. Briefing to EMC CCB, NOAA, 59 pp., https://www.emc.ncep.noaa.gov/users/Alicia.Bentley/fv3gfs/updates/EMC_CCB_FV3GFS_9-24-18.pdf.