Weather and Forecasting

Evaluation of recent NCEP operational model upgrades for cool-season precipitation forecasting over the western conterminous United States

--Manuscript Draft--

Manuscript Number:	
Full Title:	Evaluation of recent NCEP operational model upgrades for cool-season precipitation forecasting over the western conterminous United States
Article Type:	Article
Corresponding Author:	Jim Steenburgh University of Utah Salt Lake City, UT UNITED STATES
Corresponding Author's Institution:	University of Utah
First Author:	Jim Steenburgh
Order of Authors:	Jim Steenburgh
	Marcel Caron
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, limitations, and biases of these modeling systems over the western conterminous United States (CONUS), we validate and compare precipitation forecasts produced by the experimental, pre-operational 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 Forecast System (HRES). We validate using observations from Automated Surface Observing System (ASOS) stations, which are located primarily in the lowlands, and observations from Snow 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 and limitations over the western CONUS and are consistent with steady or improving NCEP model performance.
Suggested Reviewers:	

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, UT

Submitted to

Weather and Forecasting

2 September 2019

Corresponding author address: Dr. W. James Steenburgh, Department of Atmospheric Sciences, University of Utah, 135 South 1460 East Room 819, Salt Lake City, UT, 84112.
E-mail: jim.steenburgh@utah.edu

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, limitations, and biases of these modeling systems over the western conterminous United States (CONUS), we validate and compare precipitation forecasts produced by the experimental, pre-operational 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 Forecast System (HRES). We validate using observations from Automated Surface Observing System (ASOS) stations, which are located primarily in the lowlands, and observations from Snow 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 and limitations 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 and limitations 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 (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 amongst forecast systems over the western U.S. (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, pre-operational HRRRv3 and GFSv15.0 compared to their predecessor operational versions, HRRRv2 and GFSv14, respectively. The HRRRv3 upgrades include an improved planetary boundary layer (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, limitations, 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 Forecast System (HRES). A summary of the results follows in section 4.

2. Data and Methods

2.1 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 WRF model version 3.6, with physics packages and assimilation procedures described in Benjamin et al. (2016). HRRRv3 uses the WRF-ARW version 3.8, with updates to model physics, numerics, assimilated data sets, 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 developed by NOAA and run by NCEP. 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 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 (Alicia Bentley, NCEP, 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, the HRES ran with a 0.07˚ effective horizontal resolution over an octahedral reduced Gaussian grid. Parameterizations are described by Roberts et al. (2018). Operational HRES forecasts were provided by ECMWF.

2.2 Precipitation observations

Precipitation validation focuses on the CONUS west of 102.5˚ and uses observations from the Automated Surface Observing System (ASOS) and Snow Telemetry (SNOTEL) networks (Fig. 1). ASOS stations measure precipitation in 0.01-inch (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). Precipitation gauge undercatch of snowfall increases with wind speed because updrafts form over the gauge orifice, but is lower for the all-weather precipitation gauges than the standard heated tipping buckets (Greeney et al. 2005). Nevertheless, undercatch likely remains a source for measurement error during snow events (Rasmussen et al. 2012).

ASOS 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., ≥ .01 in (.254 mm); Durre et al. 2013] and received ≥ .5 inches (12.7 mm) of precipitation during the 2017/18 cool season. The resulting 277 stations (Fig. 1)—situated predominantly (but not exclusively) in lowland areas and located mainly at airports—provided 6-hour accumulated precipitation observations, which were aggregated into 24-hour totals.

SNOTEL stations are located at remote, sheltered, upland locations. Accumulated precipitation is measured hourly in 0.1-inch (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-hour 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 stations, undercatch remains a likely source of error for SNOTEL stations.

160

161 *2.3 Validation*

162 We validate model forecasts initialized between 0000 UTC 1 October 2017 and 1800 UTC
163 31 March 2018. The selection of the 2017/2018 cool season reflects the availability of forecasts
164 from all five modeling systems. To enable validation of 24-hour precipitation (hereafter daily
165 precipitation) using HRRRv2 and HRRRv3 forecasts, since the former only extends to 18 hours,
166 we combine the 6–18-hour precipitation forecasts from the 0600 UTC and 1800 UTC initialized
167 forecasts. GFSv14, GFSv15.0, and HRES validation focuses on 12–36-hour (hereafter Day 1) and
168 60–84-hour (hereafter Day 3) forecasts initialized at 0000 UTC. Periods when one or more model
169 forecasts were missing were not included, resulting in validation of 112 HRRRv2/HRRRv3 and
170 115 GFSv14/GFSv15.0/HRES daily forecasts. To compare modeled with observed precipitation,
171 we bilinearly interpolate model precipitation forecast to each station location.

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

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

182 $$HR = \frac{a}{a+c}, \quad (1)$$

false alarm ratio as

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

and equitable threat score as

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

where

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

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

3.1 Synopsis of 2017/18 cool-season precipitation

The 30-year (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, the greatest precipitation falls at stations in the Coastal, Cascade, and Olympic Mountains of the Pacific Northwest and locations in the northwest interior (Fig. 2a). For the ASOS stations used in this study, cool-season precipitation is greatest along and near the Pacific coast of northern California, Oregon, and Washington and lower in the valleys and basins of southern California and the western interior east of the Cascade–Sierra crest (Fig. 2b).

206 Integrated across all ASOS and SNOTEL stations, the 2017/18 cool-season precipitation was
207 about 40% below average. SNOTEL stations in the far north received near or slightly above
208 average precipitation, whereas stations further south received below average precipitation (Fig.
209 2c). This spatial pattern was comparatively less distinct at ASOS stations, which exhibited less
210 coherent regional patterns relative to average, especially east of the Cascade–Sierra crest (Fig.
211 2d). This likely reflects the relatively low frequency and spatial coherence of precipitation events
212 east of the Cascade–Sierra crest (Rutz et al. 2014; Touma et al. 2018), which leads to
213 undersampling at low elevation stations.
214
215 *3.2 HRRR*
216 During the 2017/18 cool season, the mean HRRRv2 bias ratio was 1.33 at ASOS stations,
217 indicating an overall wet bias (Fig. 3a). However, the bias ratio varied considerably from station
218 to station, with a standard deviation of 0.72. Forecasts for stations in northern California, Oregon,
219 and Washington west of the Cascade–Sierra crest exhibited primarily near-neutral or dry biases,
220 whereas forecasts for stations east of the Cascade-Sierra crest predominantly exhibited near-
221 neutral or wet biases. The HRRRv3 produced a similar mean bias ratio and standard deviation of
222 1.32 and 0.75, respectively, with a comparable spatial pattern of dry and wet biases at individual
223 stations (Fig. 3b). At SNOTEL stations, the mean HRRRv2 bias ratio was 0.95, with greater
224 consistency from station to station reflected in a low standard deviation (compared to forecasts for
225 ASOS stations) of 0.24 (Fig. 4a). Regions with larger dry (wet) biases include the Mogollon Rim
226 of Arizona and ranges of eastern Nevada (Big Horn Mountains of Wyoming). The HRRRv3 was
227 slightly wetter with a mean bias ratio of 1.03 and a small increase in standard deviation to 0.28
228 (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 stations, we present frequency bias for events in four bins defined by lower and upper bounds [.127–1.27 mm (.005–.05 in), 1.27–3.81 mm (.05–.15) in, 3.81–6.35 mm (.15–.25 in), and 6.35–8.89 mm (.25–.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 [0.005–0.05 in (.127–1.27 mm)], for which we use model precipitation values ≥ .127 mm (.005 in) and observed precipitation values ≥ .254 mm (.01 in). Events > 8.89 mm (.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 ≤ 6.35 mm (.25 in). HRRRv3 frequency biases were closer to neutral for events ≤ 3.81 mm (.15 in), but not significantly different from those of HRRRv3 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 (.05–.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 ≥ 1.27 mm (.05 in) and observed precipitation values ≥ 2.54 mm (.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 stations, the 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 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 precision increasing from lowland ASOS stations to upland SNOTEL stations. Low precision at the lowland ASOS stations may partially reflect undersampling. HRRRv3 is slightly wetter than HRRRv2.

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 non-events, the ETS measures skill relative to random forecasts (drawn from the observed climatological distribution). At ASOS stations, as absolute threshold increases, HRRRv2 HR decreases from 0.81 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 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.

Next, we convert absolute thresholds 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 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.

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 stations and near-neutral at upland SNOTEL stations. At both ASOS 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 to HRRRv3. We hypothesize that these differences are likely not distinguishable to operational forecasters.

298

299 *3.3 GFSv14, GFSv15.0 and HRES*

300 At ASOS stations, GFSv14 bias ratios indicate that forecasts tended to be wet, with a mean
301 bias ratio of 1.65 on Day 1 that decreases slightly to 1.57 on Day 3 (Fig. 11a and b). There are
302 large standard deviations on Day 1 (1.05) and Day 3 (1.02), which reflect large wet biases at many
303 stations. GFSv15.0 mean bias ratios are slightly higher at 1.77 on Day 1 and 1.65 on Day 3 (Fig.
304 11c and d), with comparable standard deviations. HRES forecasts were the wettest, with mean Day
305 1 and Day 3 bias ratios of 1.80 and 1.91, respectively, and comparable standard deviations (Fig.
306 11e and f). In contrast, at SNOTEL stations, mean GFSv14 Day 1 and Day 3 bias ratios are 0.99
307 and 0.97, respectively, with substantially lower standard deviations (Fig. 12a and b). GFSv15.0
308 forecasts were similar, with Day 1 and Day 3 bias ratios of 1.00 and 0.96, respectively (Fig. 12c
309 and d). HRES forecasts exhibited a weak dry bias, with mean Day 1 and Day 3 bias ratios of 0.88
310 and 0.91, respectively (Fig. 12e and f).

311 Consistent with the high bias ratios, all three models overpredicted the frequency of Day 1
312 and Day 3 precipitation events at ASOS stations for all event sizes (Fig. 13a). This problem was
313 most acute in HRES forecasts, consistent with the larger HRES wet bias. At SNOTEL stations, all
314 three models exhibited near-neutral or marginally low frequency biases on Day 1 and Day 3 for
315 all event sizes (Fig. 13b). Underprediction of event frequency was more apparent at higher
316 thresholds and increased from the GFSv15.0 to GFSv14 to HRES.

317 Bivariate histograms illustrate that GFSv14 event pairs at ASOS stations were skewed
318 above the 1:1 line, which is consistent with the aforementioned wet bias (Fig. 14a and b).
319 Furthermore, the large scatter of event pairs reflects low precision. The GFSv15.0 and HRES
320 displayed similar skewness and scatter at ASOS stations (Fig. 14c–f). At SNOTEL 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 (Fig. 15c,d). HRES, however, skewed below the 1:1 line and thus displayed slight underprediction, consistent with its weak dry bias (Fig. 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 and SNOTEL stations on Day 1 and Day 3 (Figs. 16a,b,e,f). For FAR, differences between the models are modest at ASOS 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 [$\leq$ 1.27 mm (0.05 in)] events at ASOS 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 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.

Fig. 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 and SNOTEL stations. The difference between

GFSv15.0 and GFSv14 is small on Day 1, especially at ASOS 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 stations and all thresholds at SNOTEL stations, but consistent with the ETS for absolute thresholds, 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.

In summary, all three global models produce too much and too frequent precipitation at lowland ASOS 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 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) precipitation forecasts produced by GFSv15.0 produce comparable to superior forecasts to GFSv14, although they lag HRES.

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

367 the difference was not significant at a 95% confidence level. Second, as evaluated using traditional
368 metrics, GFSv15.0 produces forecasts that are comparable to (Day 1) or superior to (Day 3)
369 GFSv14, but that still lag HRES, although the gap closes from Day 1 to Day 3. All three global
370 models (GFSv15.0, GFSv14, and HRES) produce too much and too frequent lowland
371 precipitation, but exhibit near neutral or dry biases in upland regions, with the HRES producing
372 the largest underprediction of larger upland precipitation events. These elevation-dependent biases
373 may reflect insufficient terrain representation. Superior performance of the HRES is especially
374 apparent if one verifies using event percentiles, which helps account for these biases. Operational
375 forecasters should be aware of the general biases described here, but also that there are variations
376 by location and event size.

377 These results are, however, based on a single cool season characterized by near or slightly
378 above average precipitation in the northwest CONUS and below average precipitation in the
379 southwest CONUS. Thus, precipitation events in the northwest CONUS have a strong influence
380 on overall results. Large station-by-station variations in bias ratio were identified at ASOS
381 stations, but likely reflect undersampling. Although a multi-cool-season model comparison study
382 is desirable, it is not always possible with operational modeling systems. GFSv15.0 reforecasts
383 are, however, available for three cool seasons, although for brevity we focused this paper on the
384 2017/18 cool season given that HRRRv2 and HRRRv3 were only available that cool season.

385 This study also utilized observations from the ASOS and SNOTEL networks, which
386 enables comparison of model performance in lowland and upland areas. Both station types,
387 however, likely experience undercatch, which is not accounted for here, and the quality control
388 and assessment of 24-hour precipitation amounts at SNOTEL stations is difficult and lacks data
389 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 flooding, debris flows, avalanches, and road maintenance and safety.

Acknowledgements. 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 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, and Trevor Alcott 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 No. 15, ICLR, London, Ontario, 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. *Clim. Dyn.*, **52**, 197–208, https://link.springer.com/article/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://link.springer.com/article/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. Clim.*, **26**, 380–398, https://doi.org/10.1175/JCLI-D-12-00061.1.

Chen, J.-H., X. Chen, S.-J. Lin, L. Magnusson, M. Bender, L. Zhou, and S. Rees, 2018: Tropical cyclones in GFDL fvGFS – Impacts of Dycore, physics and initial conditions. Preprints, *33rd Conference on Hurricanes and Tropical Meteorology*, Ponte Vedra, FL, Amer. Meteor. Soc., 9B4, https://ams.confex.com/ams/33HURRICANE/webprogram/Manuscript/Paper339827/9B.4_extended_abstract.pdf.

Choquet, D., P. L'Ecuyer, and C. Léger: 1999: Bootstrap confidence intervals for ratios of

expectations. *ACM Transactions on Modeling and Computer Simulation*, **9**, 326–348, DOI: 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. *Met. Apps.*, **23**, 165–81, https://doi.org/10.1002/met.1538.

Collins, W. D., and Coauthors, 2004: Description of the NCAR Community Atmosphere Model (CAM 3.0). Tech. Rep. NCAR/TN-464+STR, National Center for Atmospheric Research, Boulder, Colorado, 214 pp, http://www.cesm.ucar.edu/models/atm-cam/docs/description/description.pdf.

Craven, P. C., and Coauthors, 2018: Overview of National Blend of Models version 3.1. Part I: Capabilities and an outlook for future upgrades. Recorded presentation, *25th Conference on Probability and Statistics*, Austin, TX, Amer. Meteor. Soc., 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. Meteorol.*, **33**, 140–158, https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

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. Meteorol. Clim.*, **52**, 2377–2395, https://doi.org/10.1175/JAMC-D-13-051.1.

Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation

453 forecasts from the convection-permitting NCAR Ensemble and operational forecast
454 systems over the western United States. *Wea. Forecasting*, **33**, 739–765,
455 https://doi.org/10.1175/WAF-D-17-0144.1.
456 Greeney, C. M., M. D. Gifford, and M. L. Salyards, 2005: Winter test of production all-weather
457 precipitation accumulation gauge for ASOS 2003–2004. *Ninth Symp. on Integrated*
458 *Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface*
459 *(IOAS-AOLS)*, San Diego, CA, Amer. Meteor. Soc., 8.3,
460 https://ams.confex.com/ams/Annual2005/techprogram/paper_82895.htm.
461 Hazelton, A. T., L. M. Harris, and S.-J. Lin, 2018: Evaluation of tropical cyclone structure
462 forecasts in a high-resolution version of the multiscale GFDL fvGFS model. *Wea.*
463 *Forecasting*, **33**, 419–442, https://doi.org/10.1175/WAF-D-17-0140.1.
464 Hill, C. D., 1993: Forecast problems in the western region of the National Weather Service: An
465 overview. *Wea. Forecasting*, **8**, 158–65, https://doi.org/10.1175/1520-
466 0434(1993)008<0158:FPITWR>2.0.CO;2.
467 Horel J., M. Splitt, L. Dunn, J. Pechmann, B. White, C. Ciliberti, S. Lazarus, J. Slemmer, D. Zaff,
468 and J. Burks, 2002: Mesowest: cooperative mesonets in the western United States. *Bull.*
469 *Amer. Meteor. Soc.*, **83**, 211–225, https://doi.org/10.1175/1520-
470 0477(2002)083<0211:MCMITW>2.3.CO;2.
471 James, C. N. and R. A. Houze, Jr., 2005: Modification of precipitation by coastal orography in
472 storms crossing northern California. *Mon. Wea. Rev.*, **133**, 3110–3131,
473 https://doi.org/10.1175/MWR3019.1.
474 Lawson, J., and J. Horel, 2015: Analysis of the 1 December 2011 Wasatch downslope windstorm.
475 *Wea. Forecasting*, **30**, 115–135, https://doi.org/10.1175/WAF-D-13-00120.1.

476 Lewis, W. R., W.J. Steenburgh, T. I. Alcott, and J. J. Rutz, 2017: GEFS precipitation forecasts and
477 the implications of statistical downscaling over the western United States. *Wea.*
478 *Forecasting*, **32**, 1007–1028, https://doi.org/10.1175/WAF-D-16-0179.1.
479 Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*,
480 **21**, 289–307, https://doi.org/10.1111/j.2153-3490.1969.tb00444.x.
481 Martinaitis, S. M., S. B. Cocks, Y. Qi, and B. T. Kaney, 2015: Understanding winter precipitation
482 impacts on automated gauge observations within a real-time system. *J. Hydrometeorol.*,
483 **16**, 2345–2363, https://doi.org/10.1175/JHM-D-15-0020.1.
484 Mason, I., 2003: Binary events. *Verification: A Practitioner's Guide in Atmospheric Science*. I. T.
485 Jolliffe and D. B. Stephenson, Eds., John Wiley and Sons, 37–76.
486 McClung, T., 2014: Corrected: Global Forecast Systems (GFS) Update. Technical
487 Implementation Notice 14-46 Corrected, National Weather Service Headquarters,
488 https://www.nws.noaa.gov/om/notification/tin14-46gfs_cca.htm [Accessed 30 Aug 2019].
489 Minder, J. R., D. R. Durran, G. H. Roe, and A. M. Anders, 2008: The climatology of small-scale
490 orographic precipitation over the Olympic Mountains: Patterns and processes. *Quart. J.*
491 *Roy. Meteorol. Soc.*, **134**, 817–839, https://doi.org/10.1002/qj.258.
492 Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods:
493 Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–
494 354, https://doi.org/10.1175/2009WAF2222260.1.
495 Myrick, D. Global Forecast Systems (GFS) upgrade: Effective July 19, 2017. NWS Service
496 Change Notice 17-67. 2017. [Available online at
497 https://www.nws.noaa.gov/os/notification/scn17-67gfsupgrade.htm]
498 Myrick, D., 2018: Upgrade to the RAP and HRRR analysis and forecast system. NWS Service

Change Notice 18-58, https://www.weather.gov/media/notification/pdfs/scn18-58rap_hrrr.pdf [accessed 29 Aug 2019].

Nakanishi, M., and H. Niino, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer, *J. Meteorol. 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 from https://rapidrefresh.noaa.gov/hrrr/].

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

NWS, 2016: The Global Forecast System (GFS) – Global Spectral Model (GSM). [Accessed 29 March, 2019 from 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://www.ncbi.nlm.nih.gov/pubmed/26024836.

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. *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. Meteorol. 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, Beijing, China, http://www.wmo.int/pages/prog/www/DPFS/Meetings/ET-OWFPS_Beijing2018/linkedfiles/Doc-7-1-4_NCEP_GFS.docx [accessed 29 Aug 2019].

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. Clim.*, **31**, 8023–8036, 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, https://directives.sc.egov.usda.gov/OpenNonWebContent.aspx?content=35529.wba [accessed 29 Aug 2019].

Warner, T. T., 2004: *Desert Meteorology*. Cambridge University Press, Cambridge, UK, 595 pp.

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, https://www.emc.ncep.noaa.gov/users/Alicia.Bentley/fv3gfs/updates/EMC_CCB_FV3GFS_9-24-18.pdf [accessed 29 Aug 2019].

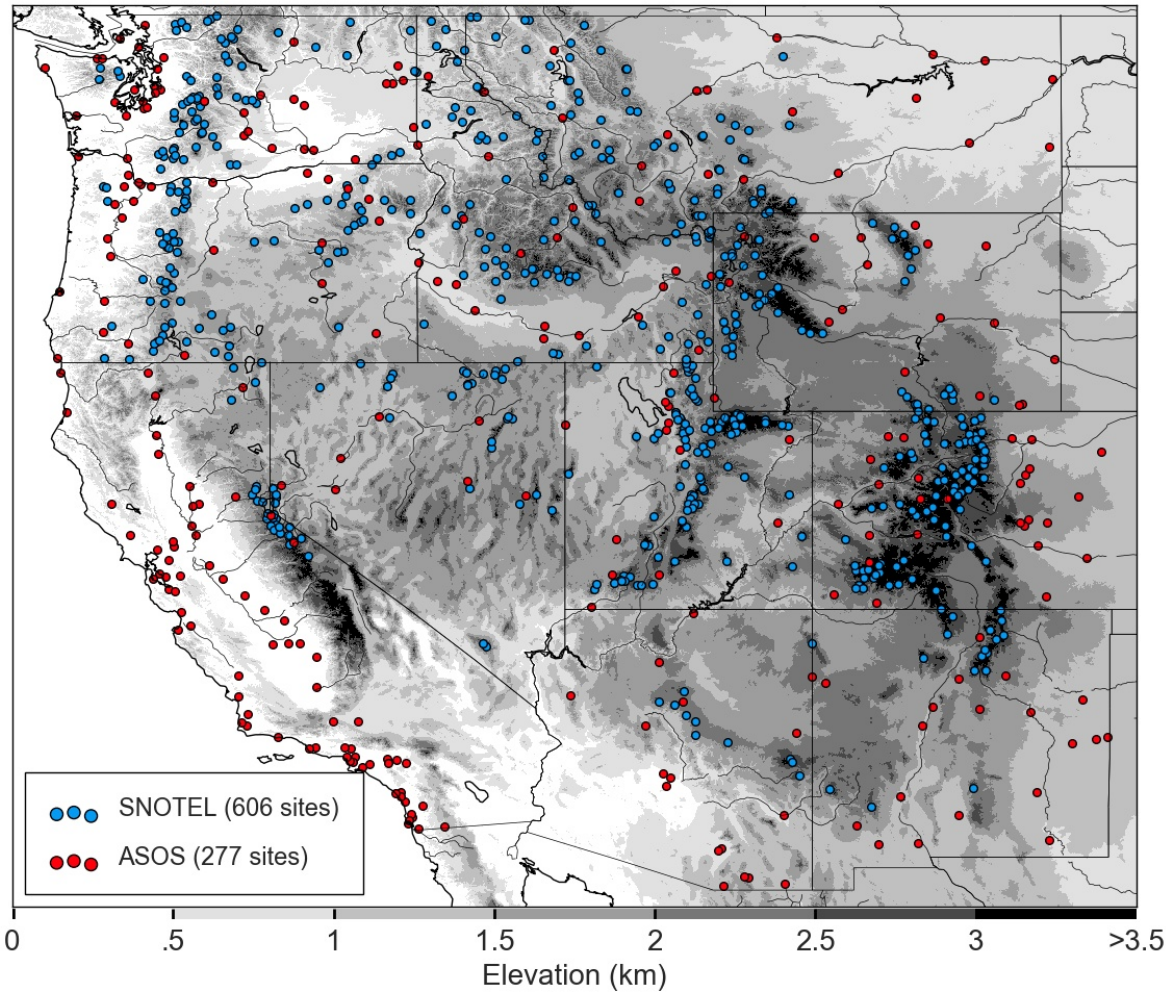
Zhou, J., and D. DeWitt, 2018: Tracking progress on NOAA's MAPP-CTB projects: Accelerating transition of research advances into improved operational capabilities. *Science and Technology Infusion Climate Bulletin*, 42nd NOAA Annual Climate Diagnostics and Prediction Workshop, Norman, OK, 76–84, https://doi.org/10.7289/V5/CDPW-NWS-42nd-2018.

Tables

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

Table 1. Contingency table used for validation

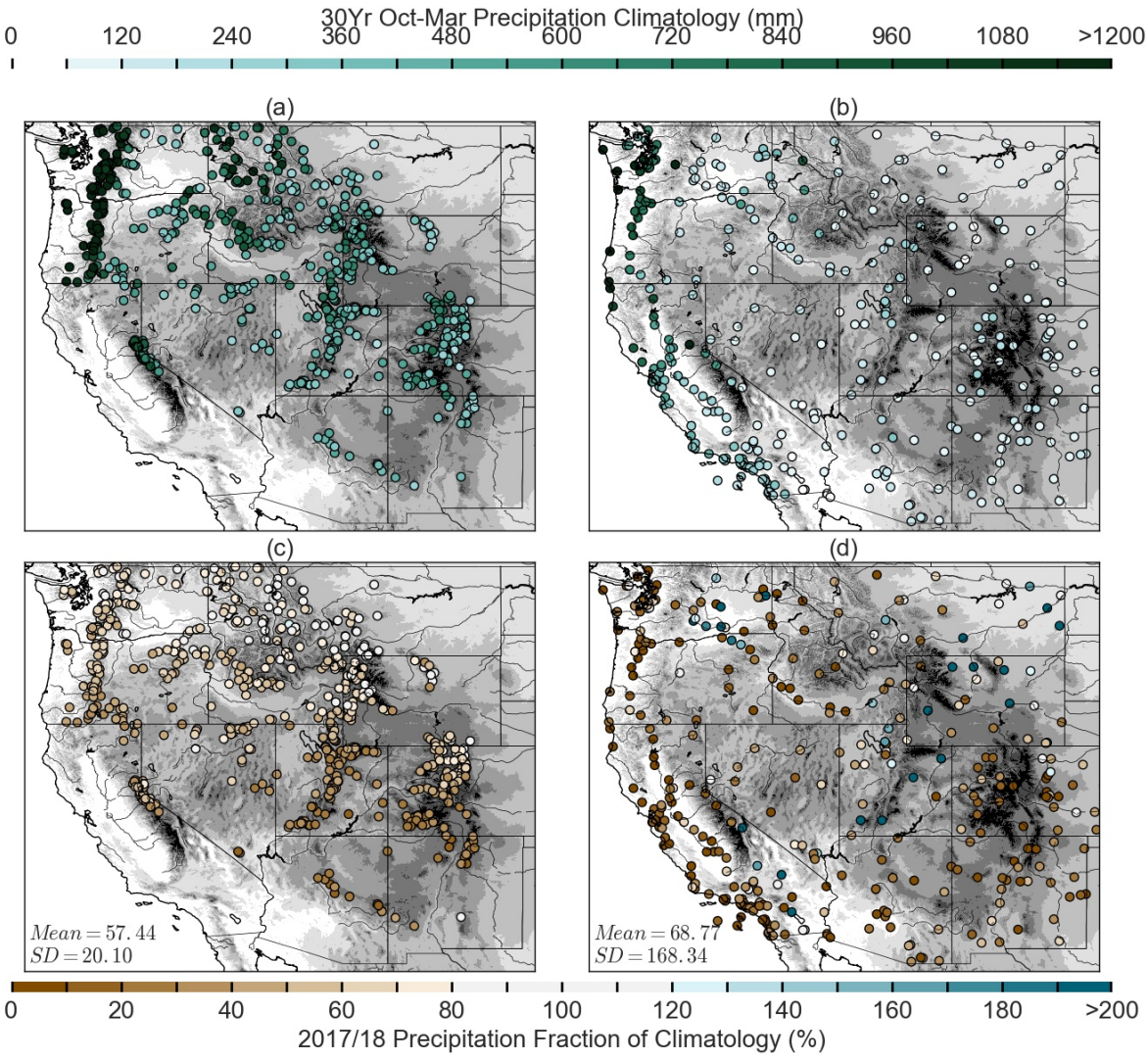
568 **Figures**



569

570 Figure 1. ASOS (red) and SNOTEL (blue) stations used for this study with 30 arc-second

571 topography (km AMSL, shaded).

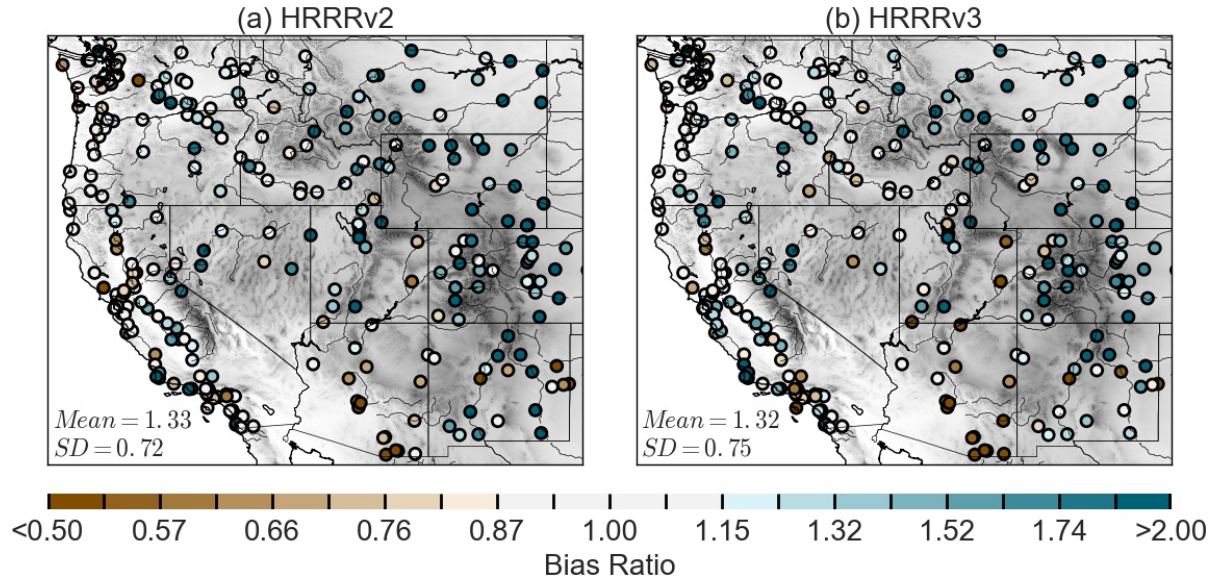


572

573 Figure 2. 30-year average accumulated cool-season precipitation at (a) SNOTEL and (b) ASOS

574 stations [based on PRISM gridded climate data (Daly et al. 1994)], and 2017/18 cool-season total

575 precipitation as a fraction of PRISM climatology at (c) SNOTEL and (d) ASOS stations.

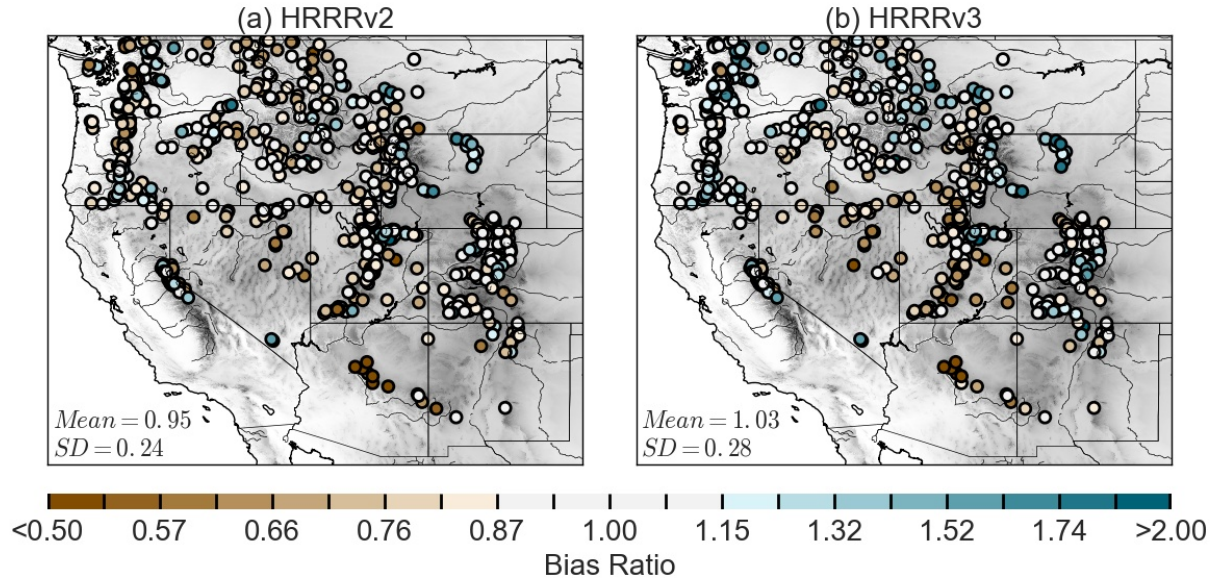


576

577 Figure 3. (a) HRRRv2 and (b) HRRRv3 bias Ratios at ASOS stations with 30 arc-second

578 topography (as in Fig. 1). Mean and standard deviation (SD) annotated.

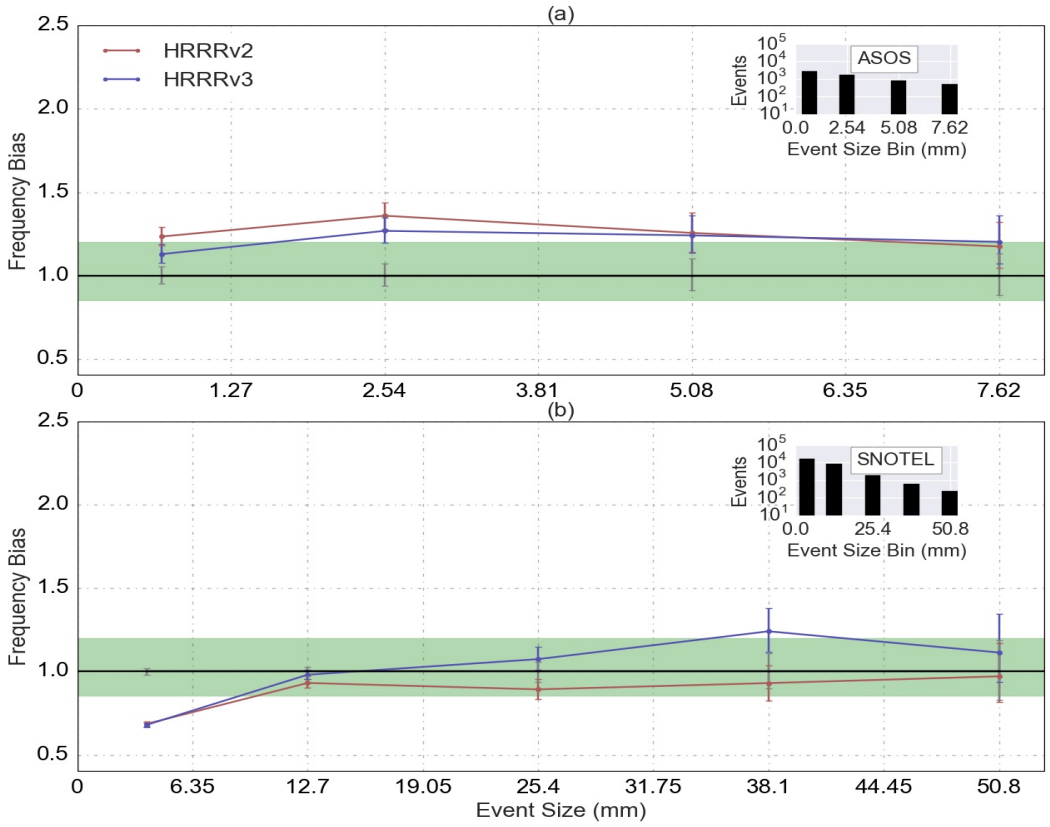
579



580

581 Figure 4. Same as Fig. 3 except for SNOTEL stations.

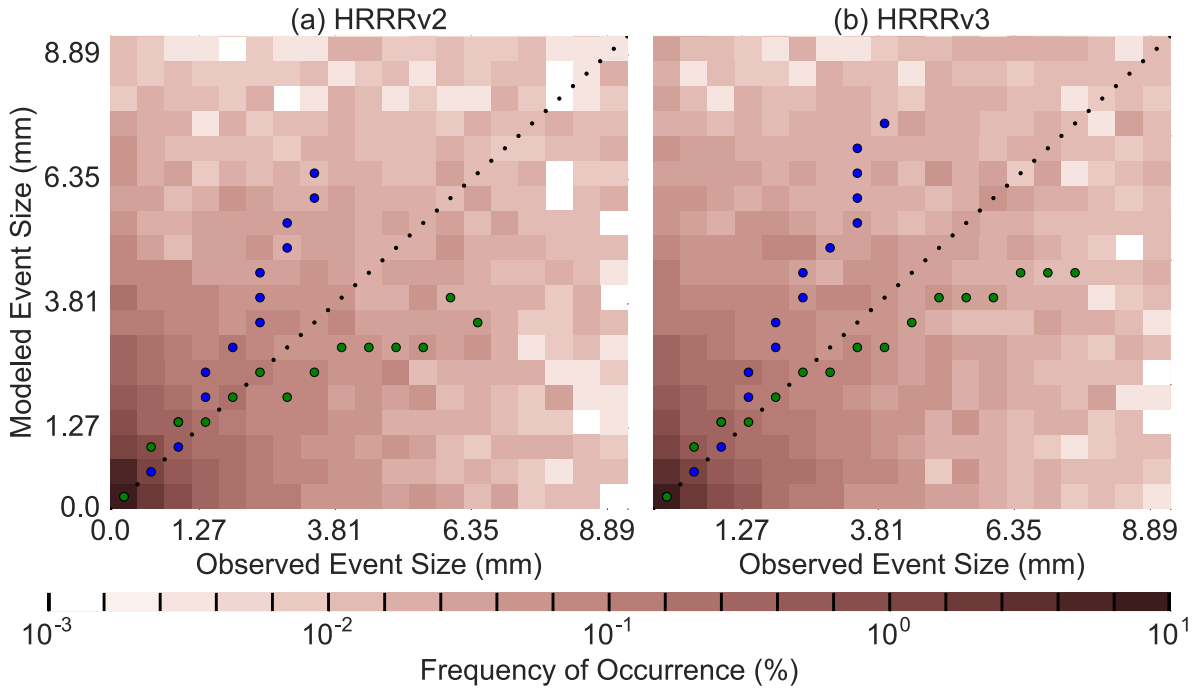
582



583

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

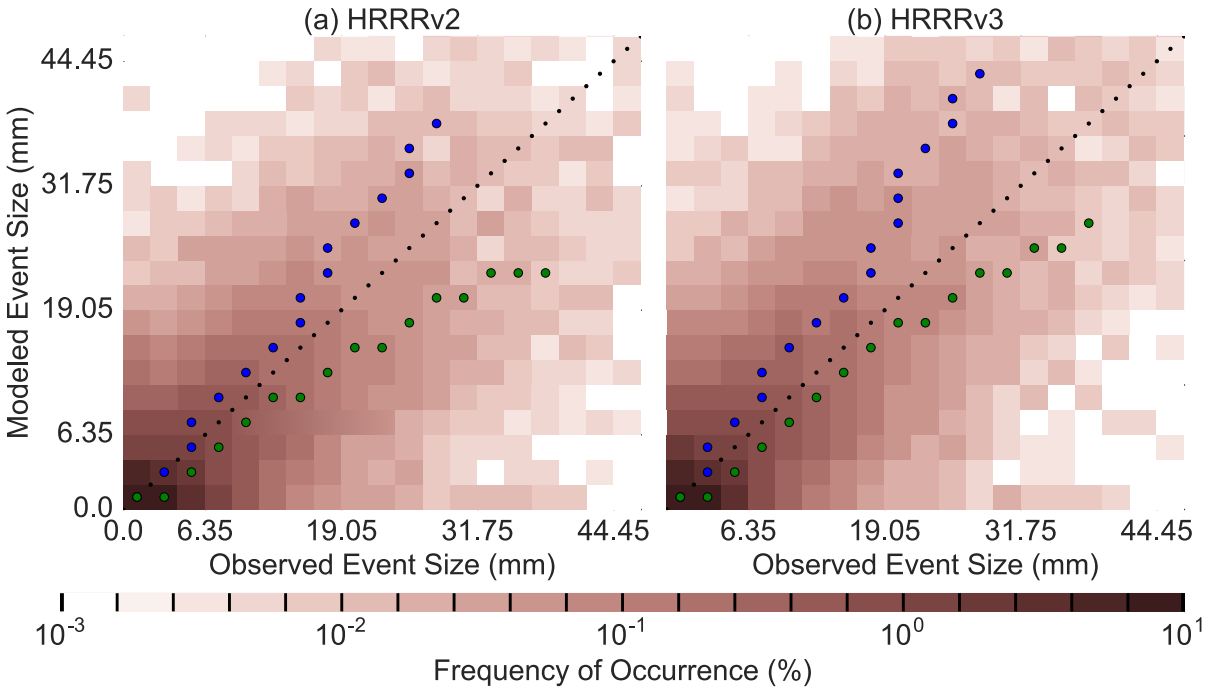
588



589

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

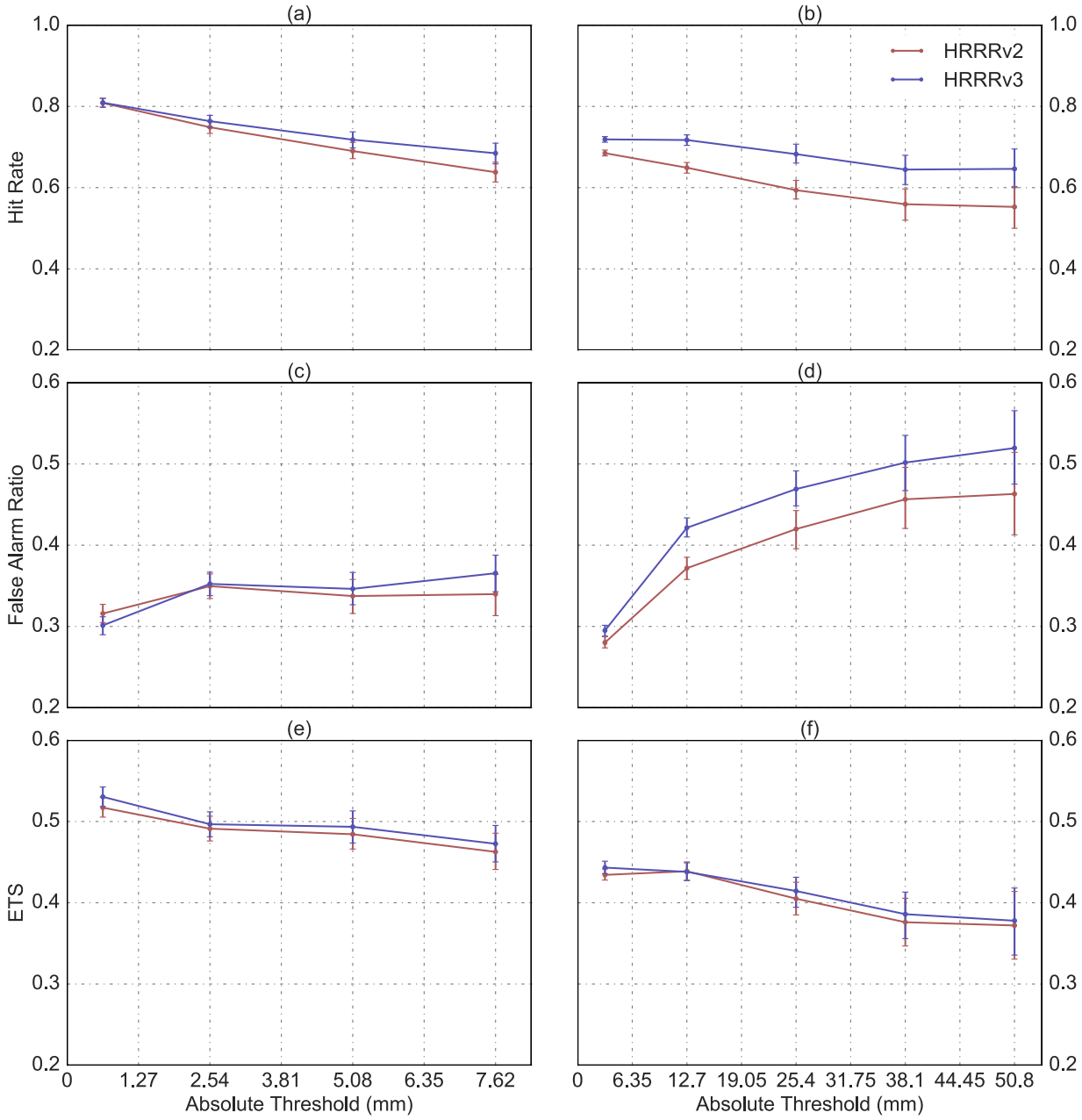
593



594

595 Figure 7. Same as Fig. 6 except for SNOTEL stations.

596



597

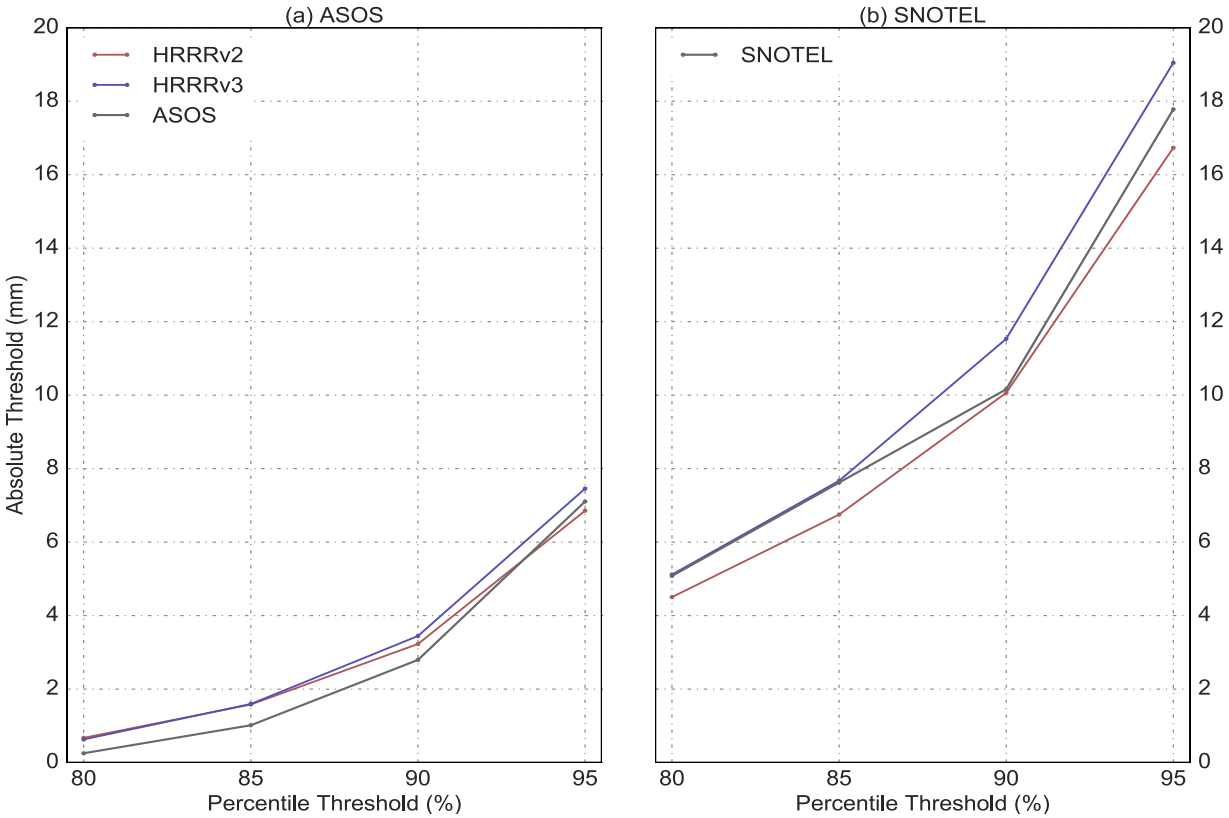
598 Figure 8. HRRRv2 (red) and HRRRv3 (blue) verification metrics as functions of absolute

599 thresholds at ASOS (a,c,e) and SNOTEL (b,d,f) stations. (a,b) Hit rate. (c,d) False Alarm Ratio.

600 (e,f) Equitable Threat Score. Whiskers display 95% confidence intervals as determined using

601 bootstrap resampling.

602

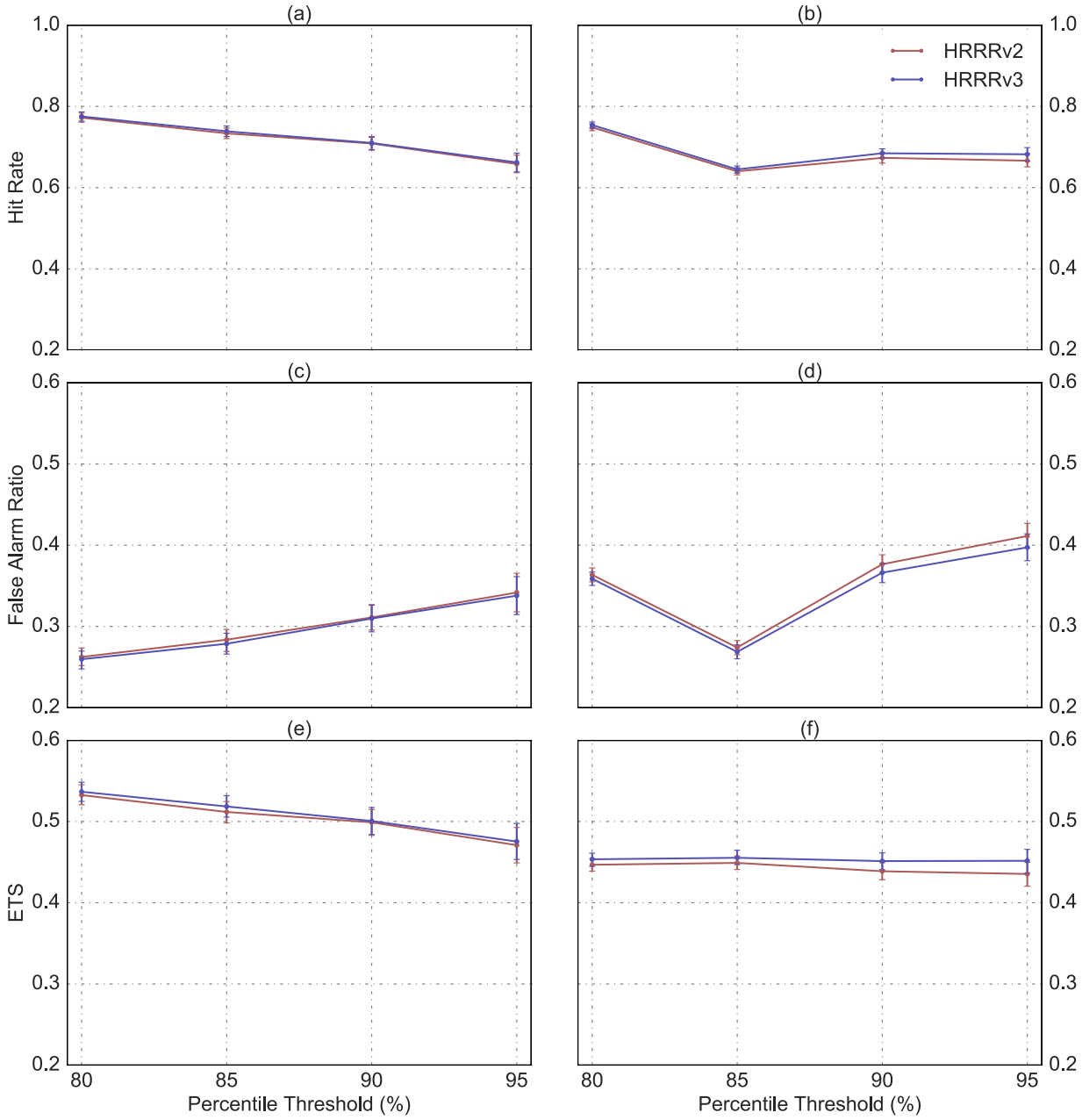


603

604 Figure 9. Observed (grey) and forecast HRRRv2 (red) and HRRRv3 (blue) absolute and

605 precipitation thresholds at (a) ASOS and (b) SNOTEL stations.

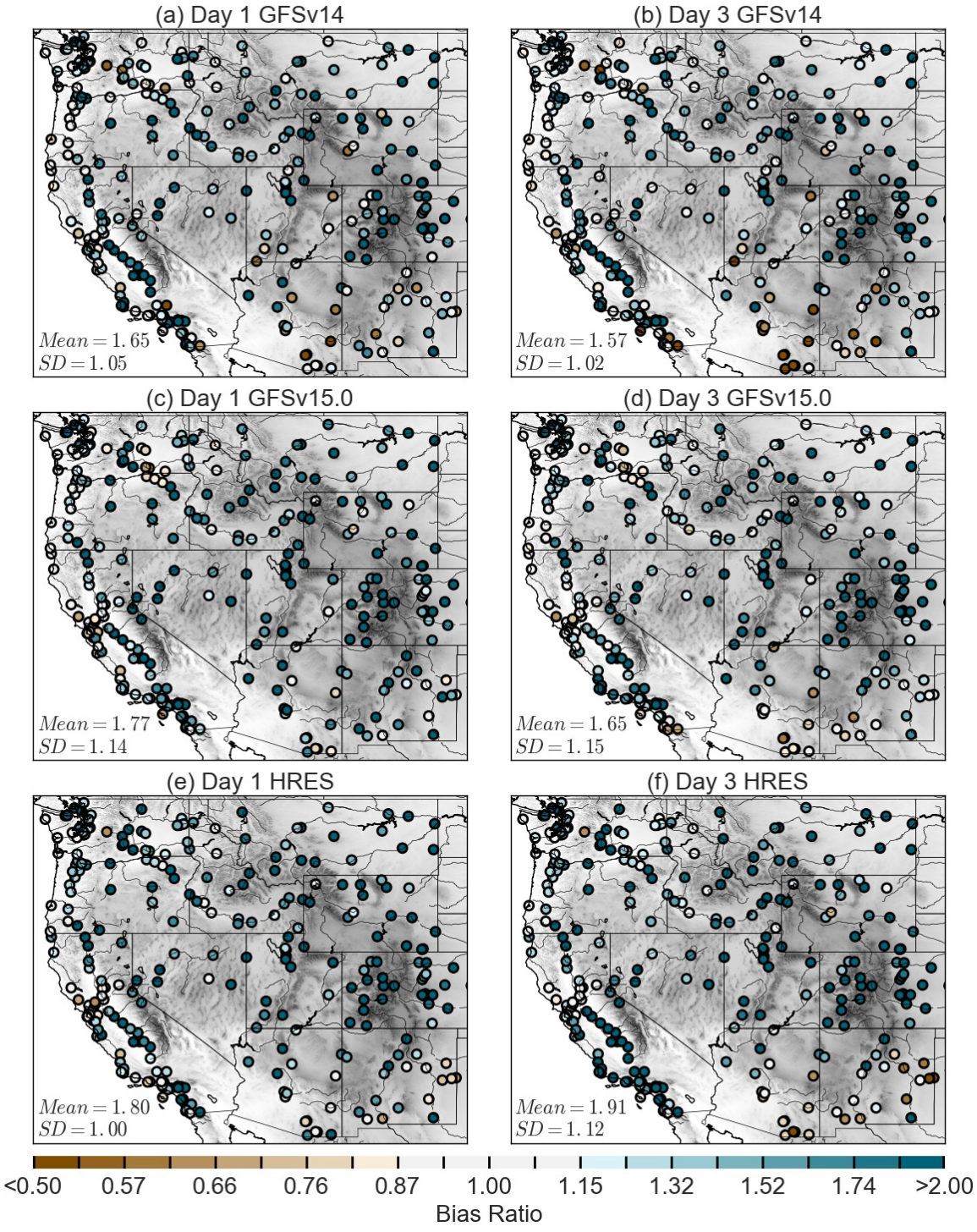
606



607

608 Figure 10. Same as Fig. 8 except for precipitation thresholds.

609

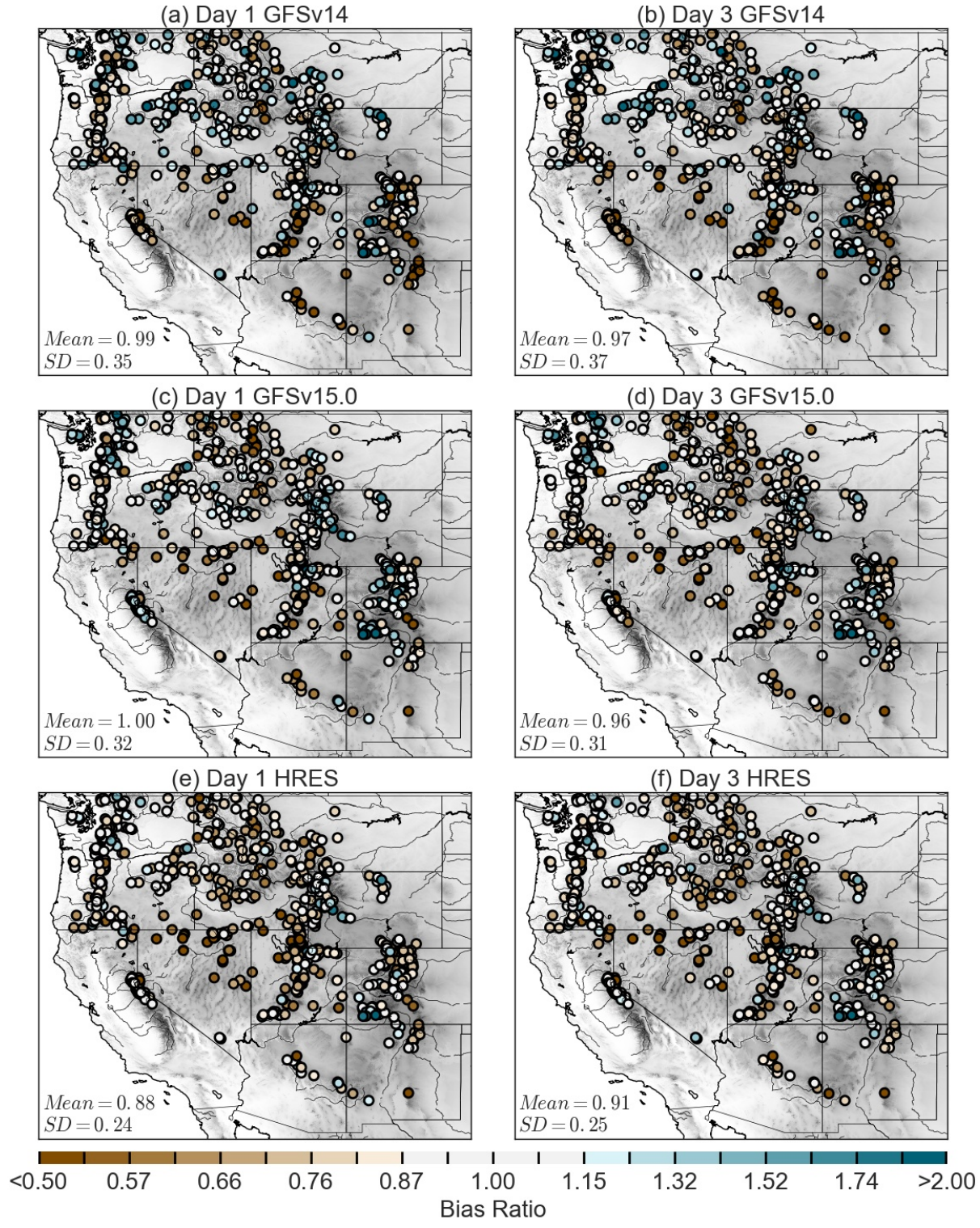


610

611 Figure 11. (a) Day 1 GFSv14, (b) Day 3 GFSv14, (c) Day 1 GFSv15.0, (d) Day 3 GFSv15.0, (e)

612 Day 1 HRES, and (f) Day 3 HRES bias ratios at ASOS stations with 30 arc-second topography

613 (as in Fig. 1). Mean and standard deviation (SD) annotated.



614

615 Figure 12. Same as Fig. 11 except for SNOTEL stations.

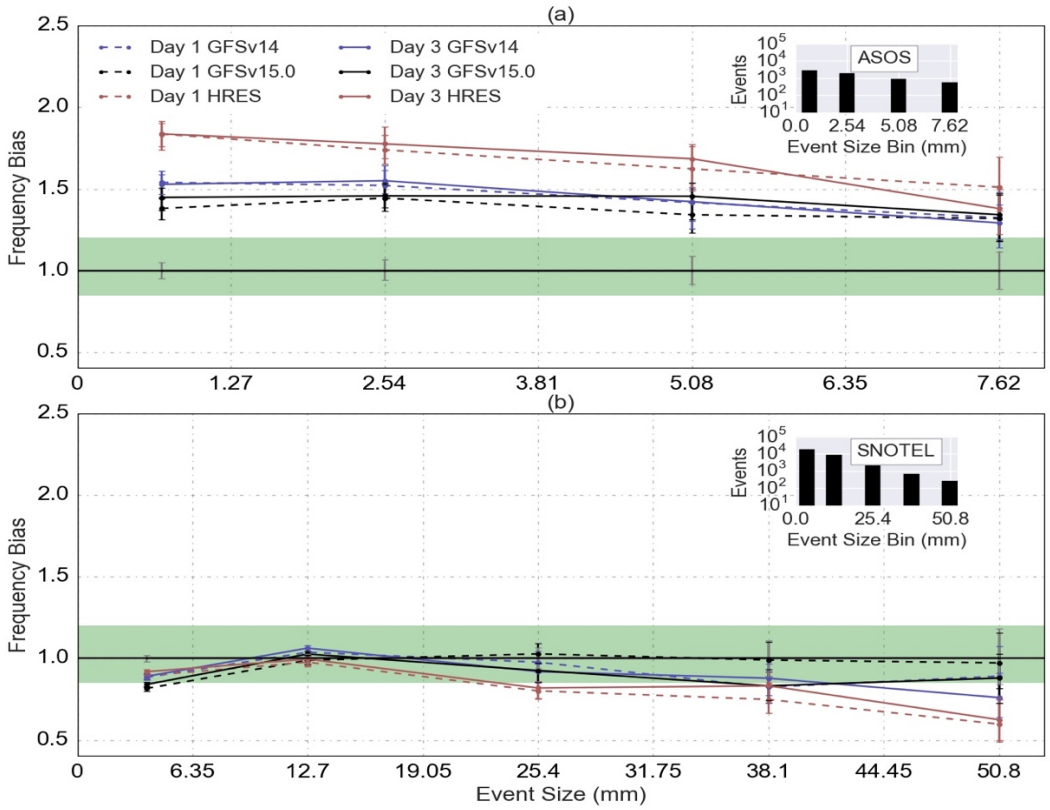
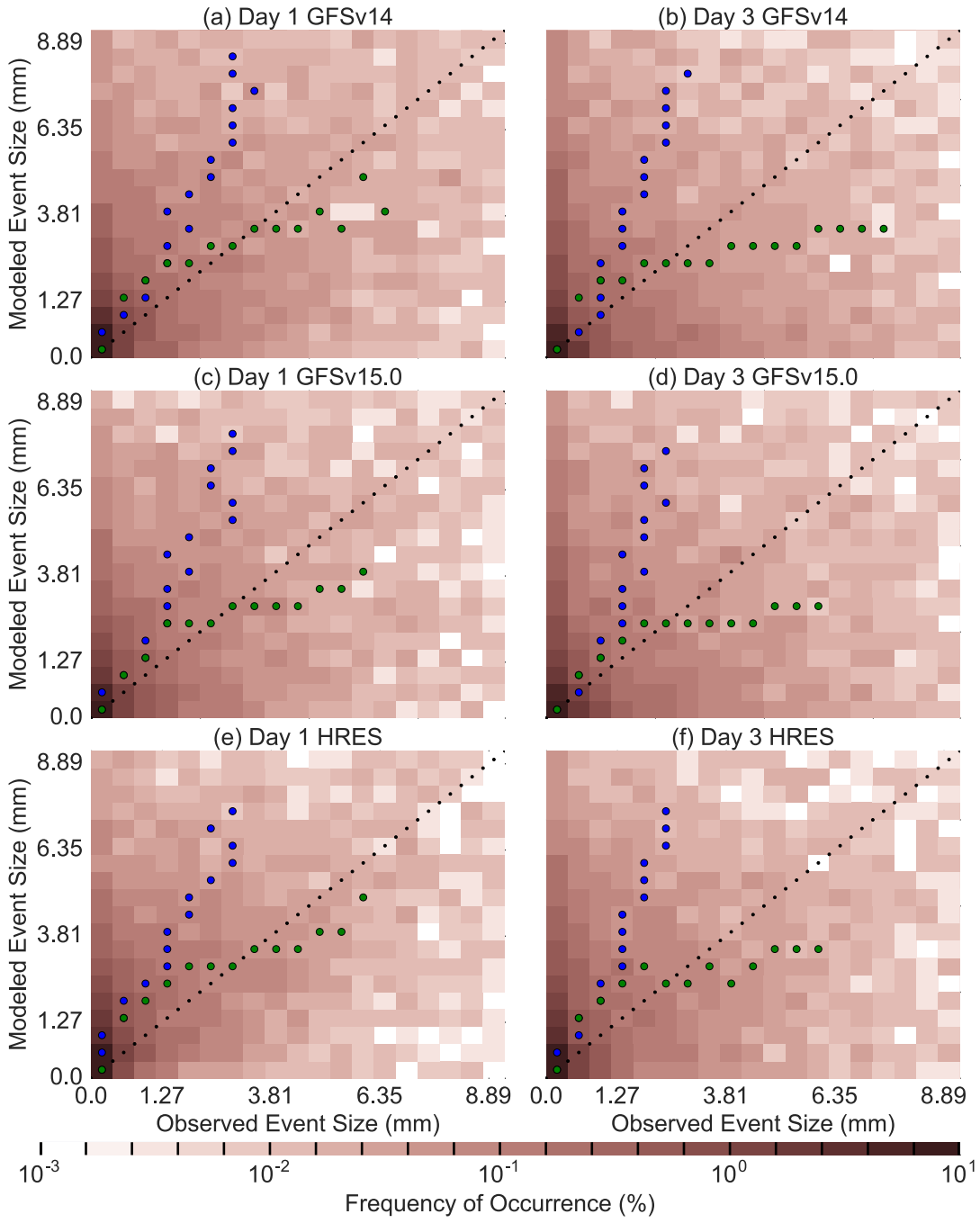


Figure 13. Day 1 (dashed) and Day 3 (solid) GFSv14 (blue), GFSv15.0 (black), and HRES (red) frequency bias as a function of event size at (a) ASOS and (b) SNOTEL stations. Number of events sampled into each bin shown in inset histograms. Green band shows 0.85–1.20 range defined as near neutral by the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



623

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

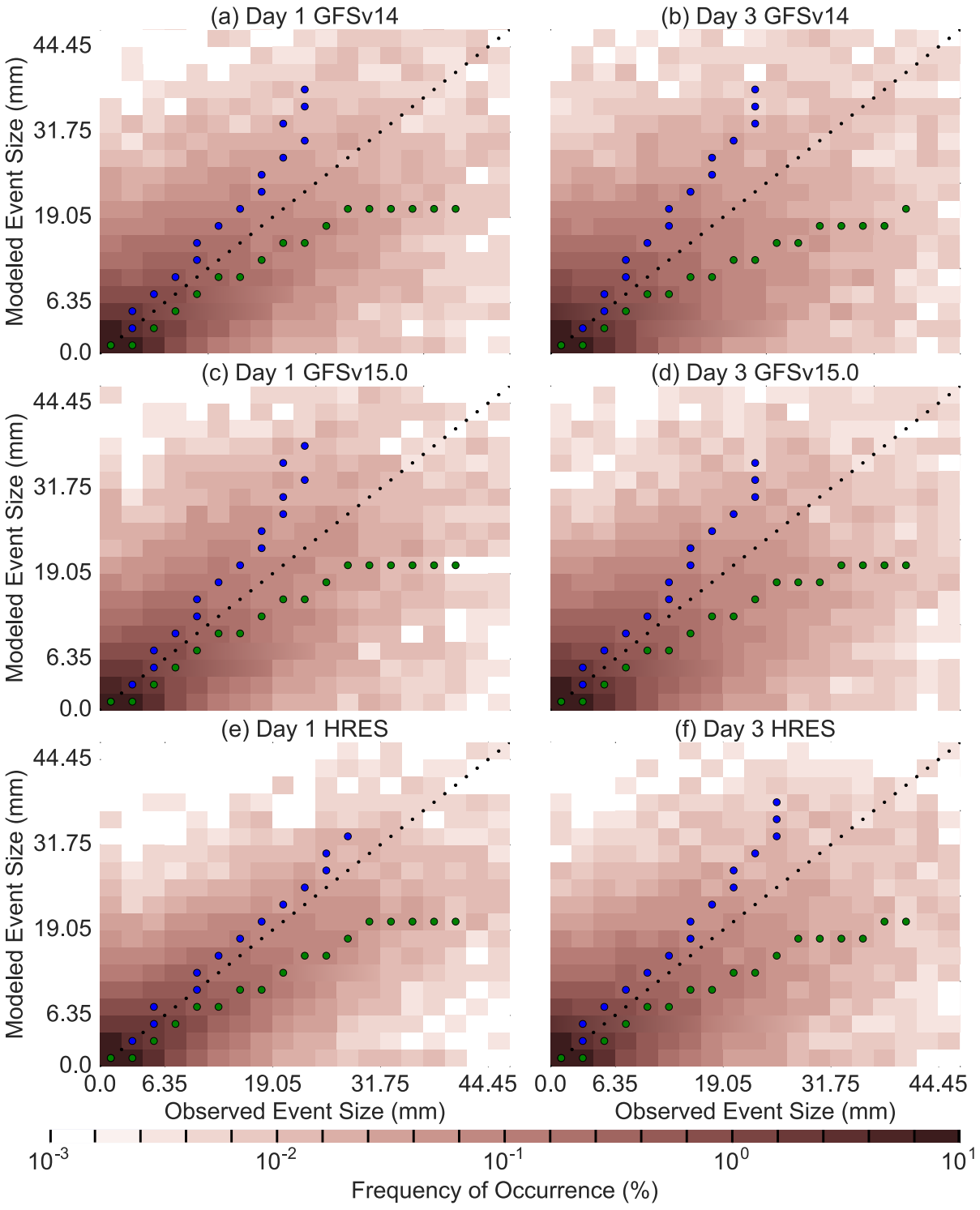
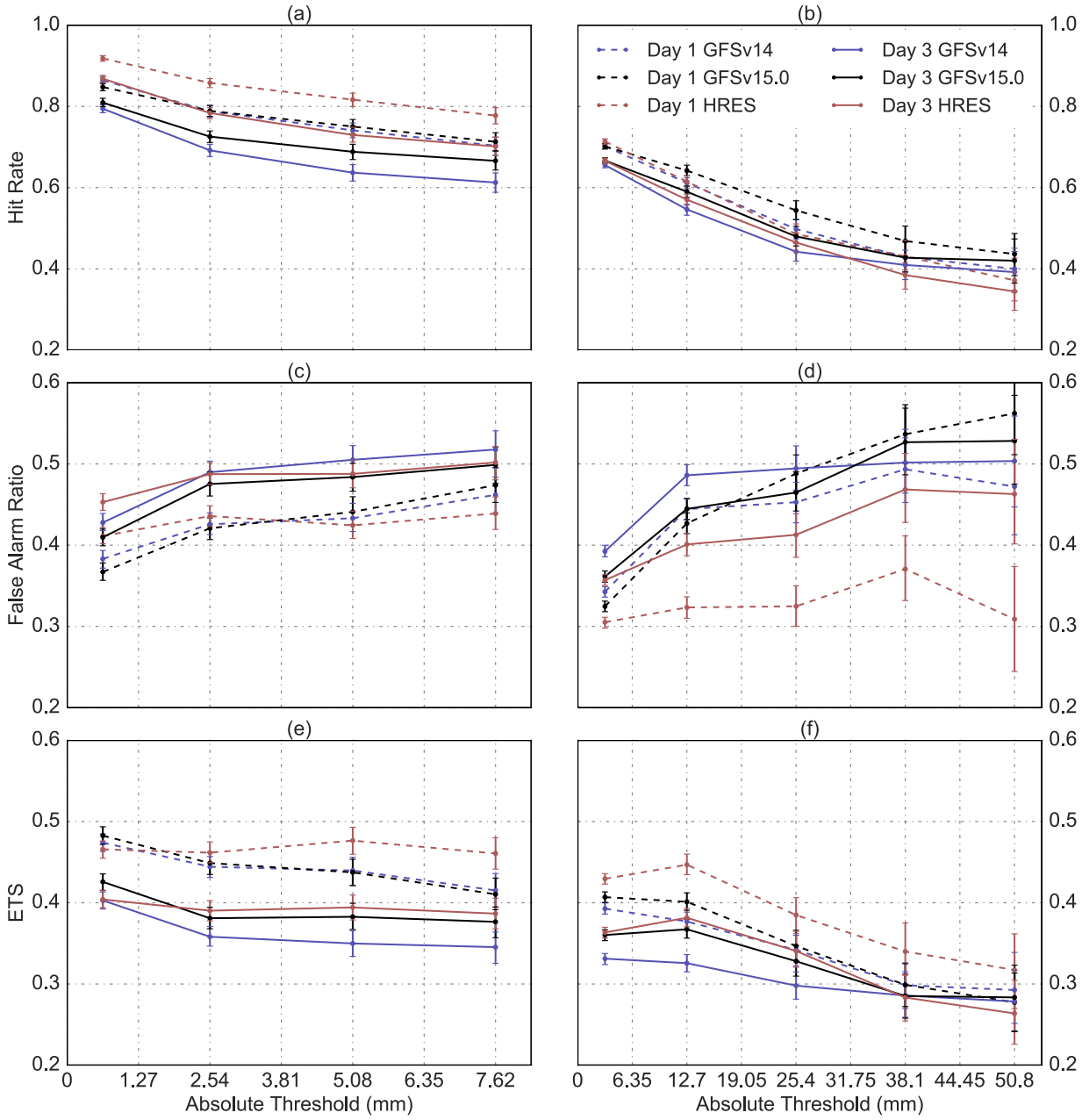


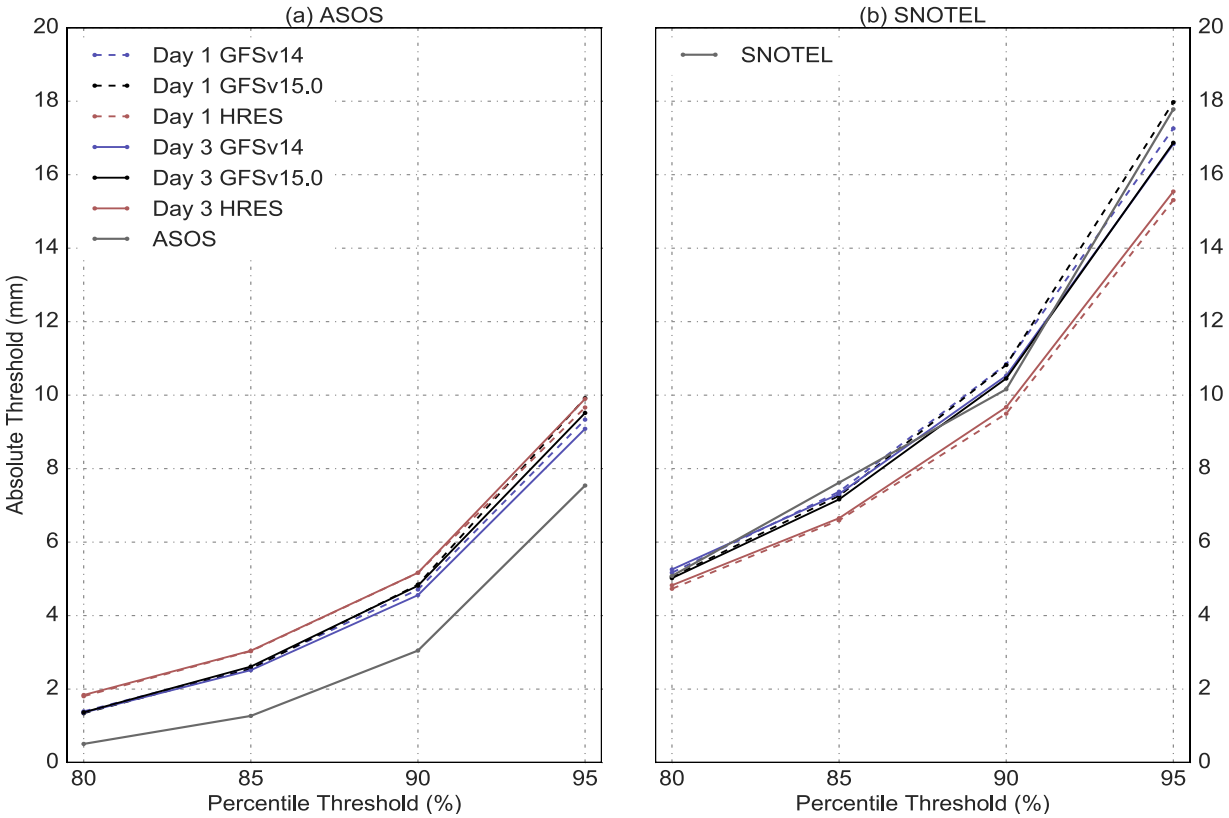
Figure 15. Same as Fig. 14 except for SNOTEL stations.



630

631 Figure 16. Day 1 (dashed) and Day 3 (solid) GFSv14 (blue), GFSv15.0 (black), and HRES (red)
632 verification metrics as functions of absolute thresholds at ASOS (a,c,e) and SNOTEL (b,d,f)
633 stations. (a,b) Hit rate. (c,d) False Alarm Ratio. (e,f) Equitable Threat Score. Whiskers display
634 95% confidence intervals as determined using bootstrap resampling.

635



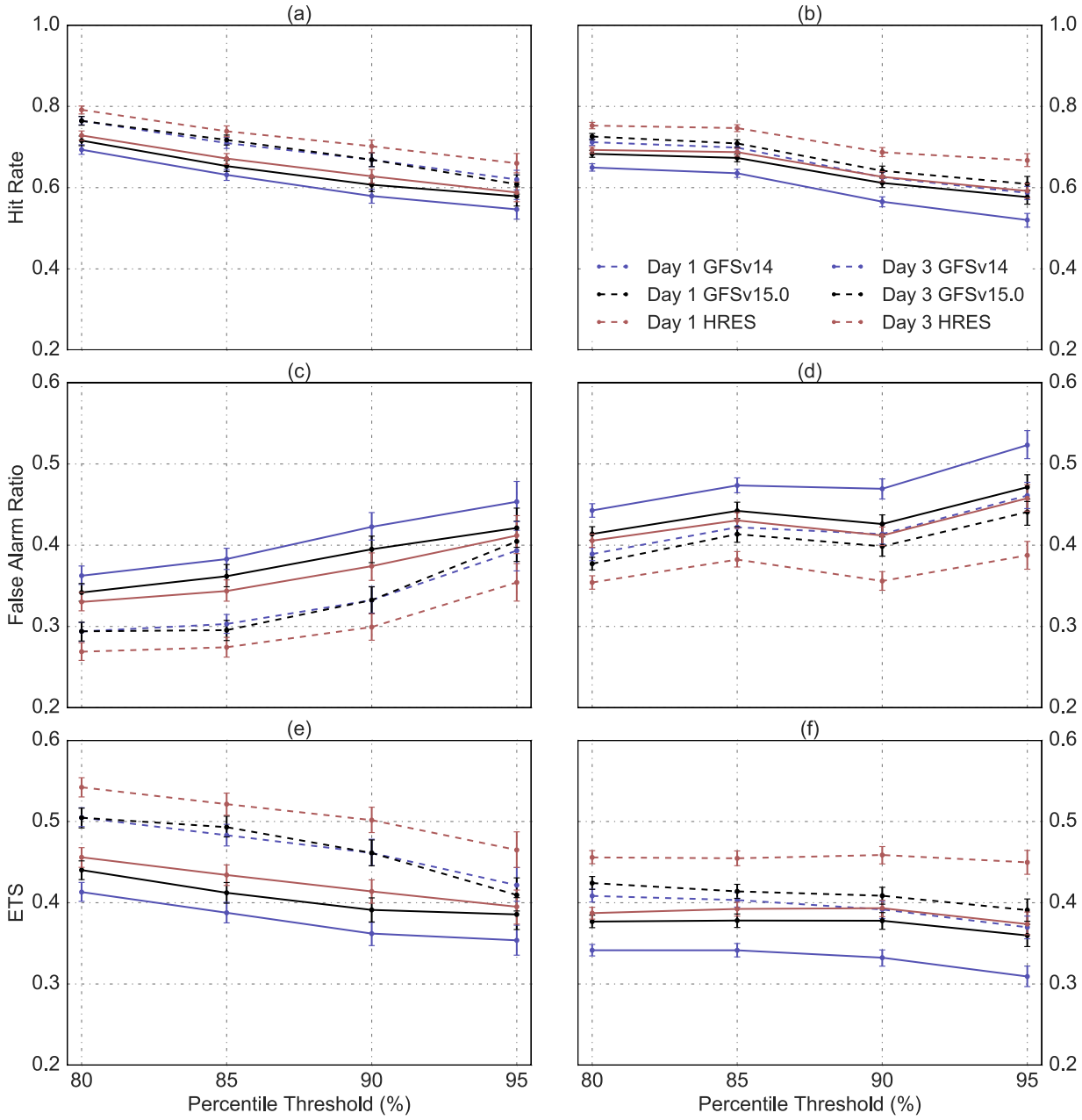
636

637 Figure 17. Observed (grey) and forecast Day 1 (dashed) and Day 3 (solid) GFSv14 (blue),

638 GFSv15.0 (black), and HRES (red) absolute and percentile precipitation thresholds at (a) ASOS

639 and (b) SNOTEL stations.

640



641

642 Figure 18. Same as Fig. 16 except for percentile thresholds.