Validation of Mountain Precipitation Forecasts from the Convection-Permitting NCAR Ensemble and Operational Forecast Systems over the Western United States

THOMAS M. GOWAN AND W. JAMES STEENBURGH

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

CRAIG S. SCHWARTZ

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 25 September 2017, in final form 31 January 2018)

ABSTRACT

Convection-permitting ensembles can capture the large spatial variability and quantify the inherent uncertainty of precipitation in areas of complex terrain; however, such systems remain largely untested over the western United States. In this study, we assess the capabilities of deterministic and probabilistic cool-season quantitative precipitation forecasts (QPFs) produced by the 10-member, convection-permitting (3-km horizontal grid spacing) NCAR Ensemble using observations collected by SNOTEL stations at mountain locations across the western United States and precipitation analyses from PRISM. We also examine the performance of operational forecast systems run by NCEP including the High Resolution Rapid Refresh (HRRR) model, the NAM forecast system with a 3-km continental United States (CONUS) nest, GFS, and the Short-Range Ensemble Forecast system (SREF). Overall, we find that higher-resolution models, such as the HRRR, NAM-3km CONUS nest, and an individual member of the NCAR Ensemble, are more deterministically skillful than coarser models, especially over the narrow interior ranges of the western United States, likely because they better resolve topography and thus better simulate orographic precipitation. The 10-member NCAR Ensemble is also more probabilistically skillful than 13-member subensembles composed of each SREF dynamical core, but less probabilistically skillful than the full 26-member SREF, as a result of insufficient spread. These results should help guide future short-range model development and inform forecasters about the capabilities and limitations of several widely used deterministic and probabilistic modeling systems over the western United States.

1. Introduction

Recent increases in computational capabilities have allowed for the development of ensemble numerical weather prediction (NWP) modeling systems with horizontal grid spacings ≤ 4 km, such that cumulus parameterizations can be omitted (Kain et al. 2008). Commonly referred to as "convection permitting" ensembles (CPEs), these modeling systems offer significant promise for improving quantitative precipitation forecasts (QPFs) and probabilistic QPFs (PQPFs) over the western United States. At present, deterministic convection-permitting models (CPMs) run operationally by the National Centers for Environmental Prediction (NCEP), such as the High Resolution Rapid Refresh (HRRR) and North American Mesoscale Forecast System 3-km continental United States (CONUS) nest (hereafter NAM-3km), provide high-resolution numerical guidance but no information concerning forecast uncertainty, except in a time-lagged sense (i.e., ensembles composed of successive model runs). In contrast, the Short-Range Ensemble Forecast system (SREF; horizontal grid spacing ~16 km) and Global Ensemble Forecast System (GEFS; effective horizontal grid spacing ~33 km) provide information on forecast uncertainty but fail to adequately resolve many key topographic features of the western United States. As a result, meteorologists employ a variety of techniques to generate QPFs and PQPFs over the western United States using deterministic CPMs (Alexander et al. 2014; Rogers et al. 2017), ad hoc ensembles composed of a collection of CPM forecasts (Alexander et al. 2011; Jirak et al. 2012, 2016),

Corresponding author: Thomas M. Gowan, tom.gowan@utah.edu

DOI: 10.1175/WAF-D-17-0144.1

coarse-resolution ensembles, and statistical downscaling approaches (Novak et al. 2014; Lewis et al. 2017).

The promise of CPEs over the western United States reflects their ability to both resolve finescale precipitation processes, including orographic effects, and estimate forecast uncertainty. The former reflects the ability of CPMs to produce precipitation forecasts with better-defined, more realistic precipitation structures than convection-parameterizing models (Mass et al. 2002; Roberts and Lean 2008; Weisman et al. 2008; Schwartz et al. 2009; Clark et al. 2016). For example, Roberts and Lean (2008) showed that forecasts of convective precipitation produced by the Met Office Unified Model (MetUM) over the United Kingdom at 1-km horizontal grid spacing without parameterized convection resulted in increased realism and skill compared to forecasts at 12-km grid spacing with parameterized convection. Similarly, Schwartz et al. (2009) found that QPFs of convection over the central United States produced by the Weather Research and Forecasting (WRF) Model at 2-km horizontal grid spacing were more detailed than those produced by the WRF at 4-km grid spacing and superior to those generated by the operational 12-km NAM. In mountainous terrain, several studies have demonstrated that decreasing horizontal grid spacing to below 4 km improves simulations of orographic precipitation (Colle et al. 2005; Garvert et al. 2005; Schwartz 2014).

Ensembles produce estimates of forecast uncertainty by executing multiple model runs, each with varied initial conditions and/or model configurations. Because of their high resolution, CPEs can assess the inherent uncertainties at convective scales, which lead to rapid error growth (Lorenz 1969), and the sensitivity of orographic precipitation to characteristics of the incident flow (Colle 2004; Roe 2005; Rotunno and Houze 2007). Using idealized simulations, Colle (2004) noted that the distribution and intensity of orographic precipitation is highly dependent on the speed of the incident flow, vertical wind shear, static stability, freezing level, and dimensions of the mountain barrier. Observational studies confirm these sensitivities and highlight the significance of low-level flow patterns (blocked or unblocked) on the distribution of orographic precipitation (Neiman et al. 2002; Stoelinga et al. 2003; Rotunno and Houze 2007; Smith et al. 2012).

Recent increases in computing capabilities in the United States have allowed for the assembling of operational, ad hoc CPEs such as the Storm Prediction Center Storm-Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012, 2016) and the HRRR Time-Lagged Ensemble (HRRR-TLE; Alexander et al. 2011), as well as the development of an experimental, but formally designed, ensemble prediction system (EPS), the NCAR Ensemble (NCAR ENS; Schwartz et al. 2015). Additionally, in Europe, several operational CPEs have been developed including the Météo-France Applications of Research to Operations at Mesoscale–Ensemble Prediction System (AROME-EPS; Bouttier et al. 2012; Vié et al. 2012), the Deutscher Wetterdienst Consortium for Small-Scale Modeling Ensemble Prediction System (COSMO-DE-EPS; Gebhardt et al. 2011), and the Met Office Global and Regional Ensemble Prediction System (MOGREPS-UK; Tennant 2015; Hagelin et al. 2017). A key difference among these CPEs is in the methods used to produce a set of forecasts. For example, the SSEO uses a multimodel, multiphysics approach (Jirak et al. 2012), whereas the HRRR-TLE simply uses a series of time-lagged forecasts (Alexander et al. 2011). The NCAR ENS, AROME-EPS, and MOGREPS-UK systems utilize ensemble data assimilation to perturb the initial conditions (Bowler et al. 2008; Vié et al. 2012; Schwartz et al. 2015), whereas nonstochastic physics perturbations are implemented in COSMO-DE-EPS (Gebhardt et al. 2011).

The majority of validation studies involving CPEs have focused on how different ensemble methods and model configurations affect their performance (e.g., Bouttier et al. 2012; Vié et al. 2012; Ben Bouallègue et al. 2013; Romine et al. 2014; Johnson and Wang 2016; Melhauser et al. 2017). Several have also investigated the ability of CPEs to forecast specific weather phenomena such as tornadoes (Gallo et al. 2016), convective initiation near the dryline (Trier et al. 2015), hurricanes (Munsell et al. 2015; Zhang and Weng 2015), and stationary convective rainbands (Barrett et al. 2016). Although limited, studies comparing the warm-season QPF performance of CPEs to convection-parameterizing ensembles have largely produced promising results (Clark et al. 2009; Duc et al. 2013; Schellander-Gorgas et al. 2017). However, work systematically intercomparing cool-season QPF performance from CPEs and convection-parameterizing ensembles is needed, as cool-season precipitation causes many hazards, such as flooding, avalanches, and traffic and air accidents. Thus, this paper evaluates the performance of cool-season QPFs produced by the 3-km, 10-member, convection-permitting NCAR Ensemble relative to several operational deterministic and probabilistic models at mountain locations throughout the western United States.

In section 2, we describe the models, datasets, and methods used in the paper, with key results and a model performance intercomparison presented in section 3. The paper concludes with a summary, including a discussion of the significance of our findings for future model development and operational forecasting over the western United States.

2. Data and methods

a. NCAR Ensemble

Described in depth by Schwartz et al. (2015), the NCAR ENS produces forecasts for the conterminous United States and consists of an analysis component run at 15-km grid spacing and a 10-member forecast component run at 3-km grid spacing. Both the analysis and forecast components use version 3.6.1 of the Advanced Research version of WRF (WRF-ARW) with 40 vertical levels and a parameterization suite that includes the Thompson microphysics scheme (Thompson et al. 2008), the Rapid Radiative Transfer Model for GCMs (RRTMG) with ozone and aerosol climatologies for long- and shortwave radiation (Mlawer et al. 1997; Iacono et al. 2008; Tegen et al. 1997), the Mellor–Yamada–Janjić (MYJ) planetary boundary layer (PBL) scheme (Mellor and Yamada 1982; Janjić 1994, 2002), and the Noah land surface model (Chen and Dudhia 2001). The analysis component also uses the Tiedtke cumulus parameterization (Tiedtke 1989). In the analysis component, an 80-member[1] continuously cycling ensemble adjustment Kalman filter (EAKF; Anderson 2001, 2003) produces analyses every 6 h (0000, 0600, 1200, and 1800 UTC). At 0000 UTC, the forecast component is initialized by interpolating 10 members of the analysis component onto a 3-km grid nested within the 15-km domain. The forecast component then produces 48-h, 10-member, 3-km forecasts. The smaller number of 3-km ensemble forecast members compared to those in the EAKF system reflects computational constraints. Nevertheless, 10 members are sufficient to produce skillful probabilistic forecasts (Clark et al. 2009, 2011; Schwartz et al. 2014). For convenience, we refer to member 1 as the control member (hereafter NCAR ENS CTL). All NCAR ENS forecasts were obtained from NCAR's Research Data Archive (RDA).

b. Operational models

We also examine the performance of several NCEP operational modeling systems including the HRRR, NAM-3km, Global Forecast System (GFS), and SREF. The SREF contains two dynamical cores, the WRF-ARW and the NCEP Nonhydrostatic Multiscale Model on the B grid (NMMB), each producing 13 ensemble members (Du et al. 2015). The control members of each core are referred to as the SREF ARW CTL and SREF NMMB CTL.

The most recent operational version of each model as of the end of the 2016/17 cool season (31 March 2017) is used for the entirety of the validation period.[2] In the case of the NAM-3km, which underwent a significant upgrade during the 2016/17 cool season (Rogers et al. 2017), parallel, preoperational runs are used prior to their operational implementation in mid-March, after which operational runs are used. HRRR and SREF forecasts were acquired from NCEP's NOAA Operational Model Archive and Distribution System (NOMADS). GFS forecasts and preoperational forecasts from the NAM-3km were provided by the NCEP Environmental Modeling Center (EMC). All modeling systems are validated using output grids at their respective horizontal grid spacing. Table 1 provides a summary of basic information for each NCEP modeling system, and Fig. 1a shows the forecast domain of each regional model.

c. Precipitation observations and analyses

Gauge-based precipitation observations from the Snow Telemetry (SNOTEL) network are used to assess the performance of QPFs and PQPFs at mountain locations. SNOTEL sites are designed to collect snowpack, precipitation, and related climatic data. There are currently over 800 sites operated and maintained by the Natural Resources Conservation Service (NRCS). SNOTEL sites are typically located in sheltered locations that receive substantial snowfall. Precipitation is measured in large storage gauges that measure hourly accumulated precipitation with a precision of 0.1 in. (~2.54 mm) using a manometer and pressure transducer (Serreze et al. 1999). Each gauge has a 30.5-cm orifice and an Alter wind shield to reduce undercatchment. Because of their sheltered locations, wind speeds at SNOTEL sites are generally less than $2\,m\,s^{-1}$ (Ikeda et al. 2010). Nevertheless, undercatchment of ~10%–15% has been shown for similar gauges under such conditions (Yang et al. 1998; Fassnacht 2004; Rasmussen et al. 2012) and likely artificially increases the model biases in our results. Such undercatch is likely more significant at sites that are windier and receive lower-density snow. Although the SNOTEL sites report hourly precipitation, we use only 24-h (1200–1200 UTC) accumulated precipitation totals to minimize the effect of artificial changes in the amount of reported precipitation as the ambient temperature fluctuates diurnally, causing the fluid in the precipitation gauges to expand and contract. Other issues that may affect SNOTEL precipitation data include transmission errors, instrument malfunction, and snow adhesion to the gauge

[1] The analysis component initially consisted of 50 members but was upgraded to 80 members in May 2016.

[2] NCEP upgraded the GFS in July 2017, so the forecasts validated here are not from the current operational version.

TABLE 1. Characteristics of modeling systems and forecasts used in this study.

Forecast system	Acronym	Modeling center	Approximate horizontal grid spacing (km)	Convective parameterization	No. of ensemble members	QPF used
NCAR Ensemble	NCAR ENS	NCAR	3	—	10	12–36 h from 0000 UTC
HRRRv2	HRRR	NCEP	3	—	—	3–15 h from 0900 and 2100 UTC
NAMv4 3-km CONUS nest	NAM-3km	NCEP	3	—	—	12–36 h from 0000 UTC
GFSv13.0.2	GFS	NCEP	13	Simplified Arakawa–Schubert	—	12–36 h from 0000 UTC
SREFv7.0	SREF	NCEP	16	Multiple	26	9–33 h from 0300 UTC

walls. Owing to these issues, we quality control the SNOTEL data following Lewis et al. (2017), resulting in data from 670 stations being available for validation. Sites that had missing or erroneous data on 20% or more of the cool-season days were removed.

We also use daily (1200–1200 UTC) precipitation analyses produced by the Parameter-Elevation Relationships on Independent Slopes Model (PRISM) Climate Group at Oregon State University (Daly et al. 1994, 2008; Di Luzio et al. 2008) to further illustrate the spatial characteristics of model biases in selected mountainous regions. PRISM analyses are produced by interpolating observational point data onto a high-resolution grid and modifying for elevation changes. The degree of modification in each grid cell is dependent on topographic aspects and the orographic effectiveness of the local terrain (Daly et al. 2008). The daily analyses are available on a ~4-km grid.

d. Verification

Although forecasts by the NCAR ENS are available beginning in April 2015, we focus on the 2016/17 cool season because of the availability of forecasts from the most recent versions of all NCEP operational models, except the GFS. Here, the 2016/17 cool season is defined as from 1 October 2016 through 31 March 2017. Each day, we validate 24-h QPFs ending at 1200 UTC on the day of interest. For example, 25 January refers to the 24-h period ending at 1200 UTC 25 January. We omitted from the study any days without an available forecast from any modeling system. Out of the 182 days during the 2016/17 cool season, 28 days are omitted, which largely reflects the unavailability of the preoperational NAM-3km runs.

For all modeling systems except the HRRR and SREF, we perform validation using the 12–36-h QPFs initialized at 0000 UTC. Because the HRRR only provides forecasts to 18 h, we merge the 3–15-h QPFs from the forecasts initialized at 0900 and 2100 UTC to obtain an equivalent 24-h QPF. We chose to begin the validation at 3 h to avoid spinup biases that have been shown to be minimal by forecast hour 3 (Bytheway and Kummerow 2015). The SREF does not run at 0000 UTC, so we use the 9–33-h QPFs from forecasts initialized at 0300 UTC. Following Lewis et al. (2017), all model QPFs are bilinearly interpolated to each SNOTEL site or PRISM grid point for calculations. Nearest-neighbor interpolation was also tested and produced nearly identical results.

In addition to deterministically verifying ensemble control members, we also evaluate the performance of the NCAR ENS, SREF ARW, and SREF NMMB ensemble mean (EM) and probability matched mean (PMM; Ebert 2001). The EM at each SNOTEL site is simply the sum of all members' QPFs at the site divided by the number of ensemble members. The PMM is calculated following the method described in Ebert (2001) with matching restricted to the western United States (see Fig. 1b for geographical area).

A thorough evaluation of QPF requires an understanding of model biases and the analysis of several statistical verification measures (Schaefer 1990; Brill 2009). Following Mason (2003), we use statistical measures based on a standard 2×2 contingency table (Table 2) to evaluate deterministic forecasts including

$$\text{hit rate} = \frac{a}{a+c} = \frac{\text{hits}}{\text{observed events}}, \tag{1}$$

$$\text{false alarm ratio} = \frac{b}{a+b} = \frac{\text{false alarms}}{\text{forecasted events}}, \tag{2}$$

and

$$\text{equitable threat score (ETS)} = \frac{a - a_{\text{ref}}}{a - a_{\text{ref}} + b + c}, \tag{3}$$

where

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

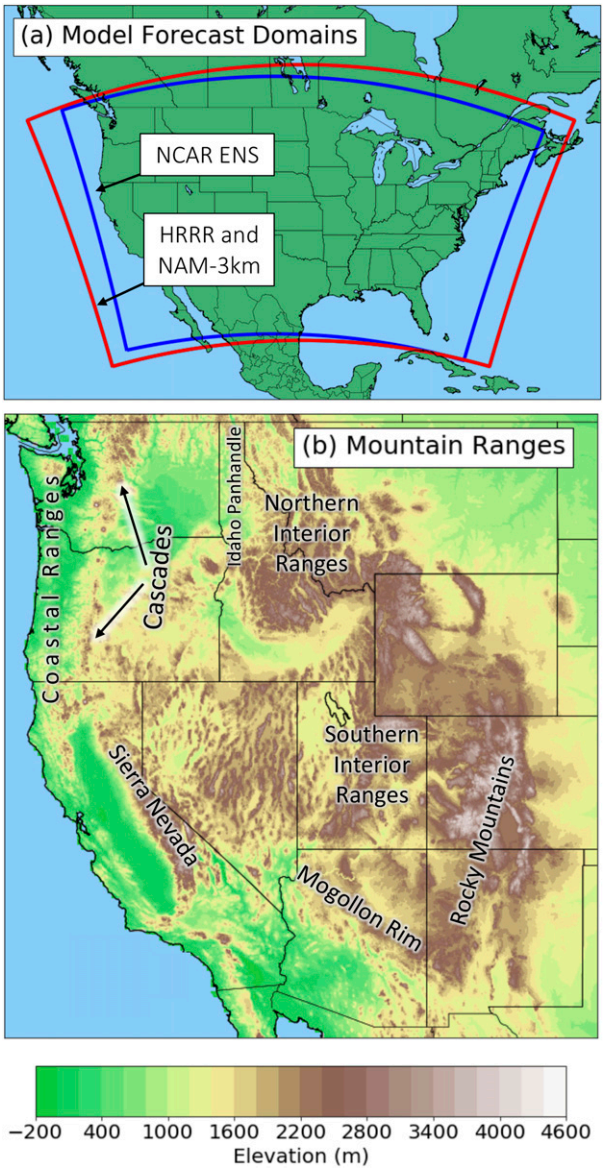

FIG. 1. (a) NCAR ENS, HRRR, and NAM-3km forecast domains. The SREF forecast domain covers all of North America, and the GFS forecast domain covers the entire globe. (b) Geographic terms referenced in text and 30-arc-s topography (m MSL, color scale at bottom).

a is the number of hits, b is the number of false alarms, c is the number of misses, d is the number of correct rejections, and n is the total number of forecast–observation pairs. Hit rate measures the fraction of observed events correctly forecasted, false alarm ratio expresses the fraction of forecasted events that were false alarms, and ETS measures the fraction of observed and/or forecasted events that were correctly forecasted, adjusted for the frequency of hits expected by chance (climatology). While modern, convective-scale verification measures including "neighborhood" approaches have been developed (e.g., Ebert 2008), we use the point-based ETS because cool-season precipitation in mountainous regions is strongly tied to terrain. Issues would arise using neighborhood approaches because of the dramatic changes in precipitation climatology over small spatial scales that exist in mountainous regions. Although varying climatological event frequencies among sites can affect ETS (Hamill and Juras 2006), we use the traditional ETS since Lewis et al. (2017) found that a weighted average ETS for 10 subgroups of SNOTEL sites was similar to the traditional ETS.

TABLE 2. Contingency table used for validation.

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

We determine the quality of probabilistic forecasts from ensembles by computing their reliability and resolution, which are defined as

$$\text{reliability} = \frac{1}{N}\sum_{k=1}^{K} n_k(f_k - \overline{o}_k)^2 \quad \text{and} \tag{5}$$

$$\text{resolution} = \frac{1}{N}\sum_{k=1}^{K} n_k(\overline{o}_k - \overline{o})^2, \tag{6}$$

where N is the total number of forecasts, K is the total number of unique probabilistic forecast bins, $\overline{o}$ is the observed climatological frequency of the occurrence of the event, n_k is the number of forecasts in the kth bin, and $\overline{o}_k$ is the observed frequency of the occurrence of the event given forecasts of probability f_k. Reliability assesses the statistical consistency between predicted probabilities and observed relative frequencies, whereas resolution measures the ability of an ensemble to distinguish when the event of interest occurs with lower or higher frequency than climatology. We also calculate the Brier score (BS) for each ensemble, which measures the mean squared probability error and is given by

$$\text{BS} = \text{reliability} - \text{resolution} + \text{uncertainty}, \tag{7}$$

where

$$\text{uncertainty} = \overline{o}(1 - \overline{o}). \tag{8}$$

Additionally, we measure the skill of the ensemble by computing the Brier skill score (BSS; Brier 1950; Murphy 1973; Wilks 2011), which is defined as

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS_{Cl}}}, \tag{9}$$

where $\mathrm{BS_{Cl}}$ is the BS of climatology and equal to Eq. (8). Good ensemble performance is indicated by lower values of BS and reliability and higher values of BSS and resolution. We use attributes diagrams to visually assess these statistical measures and evaluate other ensemble characteristics (Toth et al. 2003). Consistency resampling (Brocker and Smith 2007) and bootstrap resampling (Hamill 1999; Efron and Tibshirani 1993) are employed to produce 5% and 95% consistency bars and confidence intervals, respectively, for the attributes diagrams, which involve resampling 1000 times and choosing N samples with replacement.

All of these measures require that the event of interest be dichotomous (yes/no). Therefore, we apply a threshold to each event, which we define as the total accumulated observed or forecast precipitation during a 24-h (1200–1200 UTC) period (including 24-h periods with no precipitation). In addition to using absolute event thresholds (e.g., 15, 20, 25 mm, etc.), we use event percentile thresholds (e.g., 75th, 80th, 85th percentile, etc.). Following Roberts and Lean (2008) and Dey et al. (2014), we compute the distribution of events observed at SNOTEL sites and forecasted by each deterministic model and ensemble member to determine percentile thresholds for the observed and forecast events. Because we compare percentile thresholds from observed and forecast events, the absolute thresholds corresponding to a given percentile threshold for the observations and forecasts can differ. For example, the 95th percentile, which represents the top 5% of events, may be 35 mm for a certain model and 25 mm for SNOTEL observations. This method implicitly removes bias, allowing for an assessment of the placement of precipitation within the context of each model's climatology and reduces sampling issues resulting from differing observed and forecast precipitation climatologies across the western United States.

3. Results

a. *Observed and forecast cool-season precipitation characteristics*

1) Synopsis of 2016/17 cool-season precipitation

Significant spatial variations in precipitation existed across the western United States during the 2016/17 cool season. The cool season was generally wetter than average across all of the western United States, except for portions of Colorado, southern Utah, Arizona, and New Mexico, where precipitation was average to slightly below average (not shown). At upper elevations, mean daily precipitation ranged from >16 mm in the Cascades and coastal ranges of the Pacific Northwest to <3 mm in parts of the Rocky Mountains of Colorado and New Mexico, as well as other climatologically dry ranges of the western U.S. interior (Figs. 2a,b; see Fig. 1b for geographic references). Measureable precipitation (≥2.54 mm)[3] occurred on ~70%–80% of days in the Cascades and coastal ranges of the Pacific Northwest, ~45%–70% of days in the Sierra Nevada and northern interior ranges, and ~25%–45% of days in the southern interior ranges (Figs. 3a,b). The magnitudes of the 85th and 95th percentile events were generally greatest in the Cascades, coastal ranges from northern California to Washington, and Sierra Nevada, and decreased toward the interior ranges (Figs. 4a–d). SNOTEL sites with relatively large 85th and 95th percentile events in the interior were found in the Idaho panhandle and the Mogollon Rim of Arizona (Figs. 4a,c), regions that receive relatively large fractions of their climatological cool-season precipitation from inland-penetrating atmospheric rivers (Rutz et al. 2014, 2015).

2) Model biases

The ratio of forecast to observed mean daily precipitation (i.e., the bias ratio) identifies SNOTEL site locations where a model over- (bias ratio > 1) or under- (bias ratio < 1) predicts the total observed cool-season precipitation. Given undercatch and observational uncertainty, we consider bias ratios of 0.85–1.2 to be reflective of a near-neutral bias. For ensembles, we focus on the control member of each dynamical core. The NCAR ENS has one (NCAR ENS CTL) and the SREF two control members (SREF ARW CTL and SREF NMMB CTL). Other members in each core exhibit similar bias ratios as their respective control runs, as will be shown in section 3c. At SNOTEL sites, the NCAR ENS CTL produces mean bias ratios ~ 1 with relatively low standard deviations of bias ratios over all SNOTEL sites, indicating its ability to accurately produce total cool-season precipitation at mountain locations (Fig. 5a). Aside from a dry bias at SNOTEL sites in Idaho and northwest Montana, the HRRR exhibits bias ratios similar to the NCAR ENS CTL (Fig. 5b). The NAM-3km exhibits a large mean bias ratio of 1.319, indicative of a substantial wet bias (Fig. 5c). Although the GFS and SREF ARW CTL also

[3] The precision of the precipitation gauges at SNOTEL sites is 0.1 in. (2.54 mm). Hence, the minimum amount of precipitation they can record is 2.54 mm.

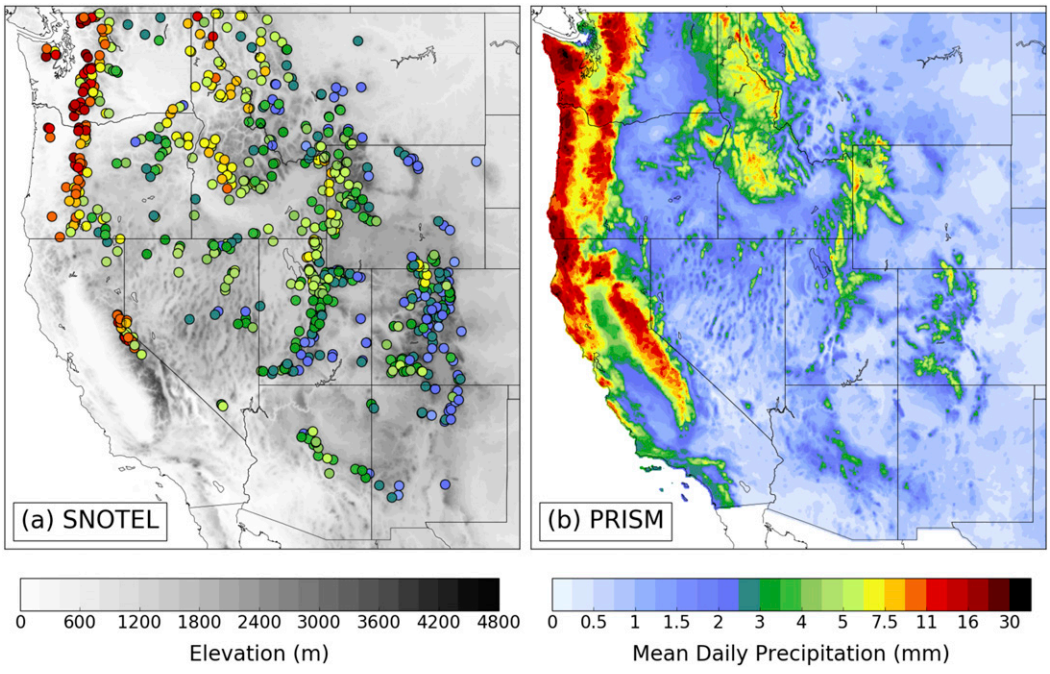

FIG. 2. (a) Mean daily precipitation at SNOTEL sites (mm; color scale at bottom right) and 30-arc-s topography (m MSL; gray-shaded scale at bottom left). (b) Mean daily precipitation from PRISM analyses [mm; color shaded as in (a)].

produce mean bias ratios of ~1, relatively high standard deviations (0.397 and 0.434, respectively) reflect sizeable dry or wet biases at individual SNOTEL sites (Figs. 5d,e). In contrast, the SREF NMMB CTL has a significant dry bias, especially in southern Utah and Colorado (Fig. 5f).

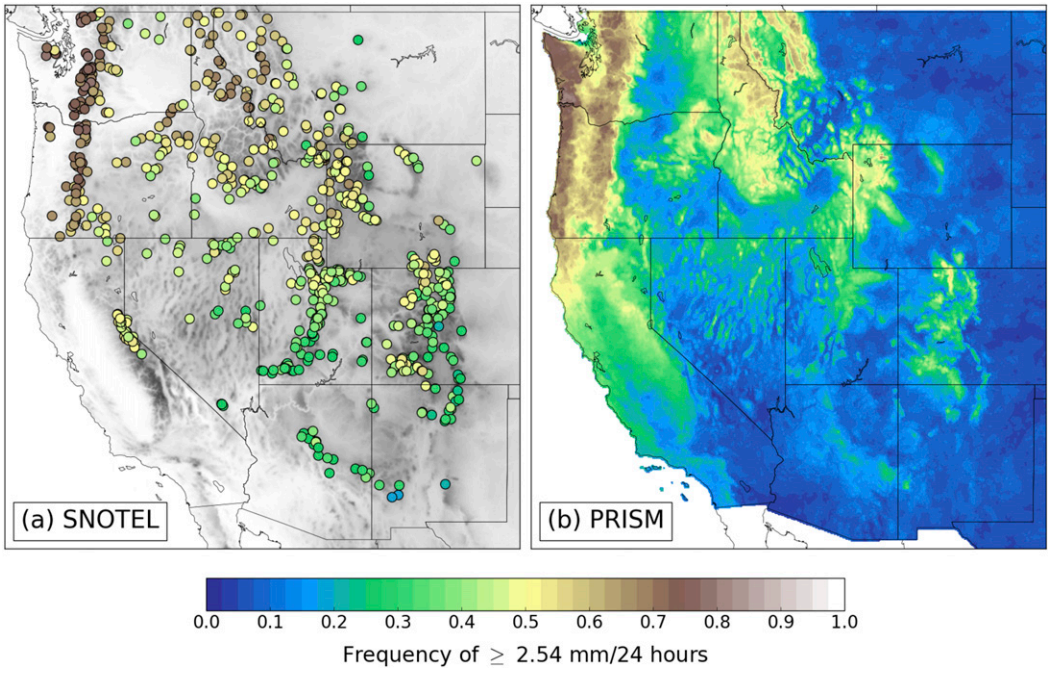

FIG. 3. (a) Frequency of precipitation events (≥2.54 mm) from SNOTEL observations (color scale at bottom) and 30-arc-s topography (as in Fig. 2a). (b) As in (a), but from PRISM analyses.

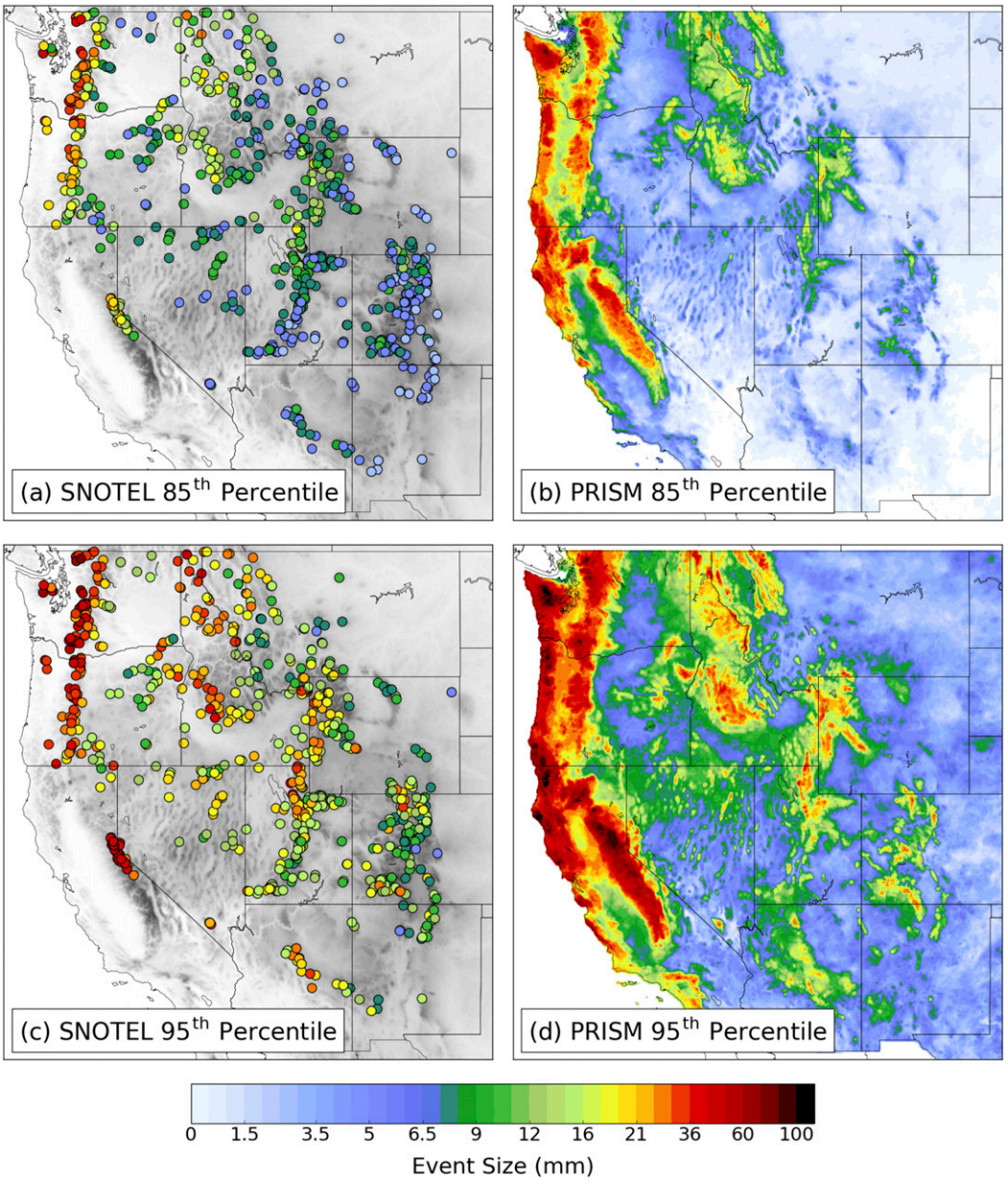

FIG. 4. (a) Magnitude of 85th percentile events at SNOTEL sites (mm; color scale at bottom) and 30-arc-s topography (as in Fig. 2a). (b) As in (a), but from PRISM analyses. (c),(d) As in (a),(b), but for 95th percentile events.

Following Lewis et al. (2017), we divide the SNOTEL sites into two regions, Pacific ranges and interior ranges, that feature highly differentiated climatologies and terrain characteristics (Fig. 6). Intermediate stations are not presented for brevity. Time series of accumulated precipitation averaged over all SNOTEL sites in each region provide information regarding regional model biases (Fig. 7). The NCAR ENS CTL generated ~112% of the total observed precipitation over the Pacific ranges and about as much precipitation as observed by SNOTEL sites over the interior ranges. The HRRR produced only ~86% of the total observed precipitation in the interior ranges, reflective of a dry bias, but agreed more closely with observations in the Pacific ranges. Total precipitation produced by the GFS was close to observed in both regions. The NAM-3km produced excessive precipitation in both regions, especially over the interior ranges where it produced ~130% of the total observed precipitation. The SREF ARW CTL's total predicted precipitation was slightly greater than

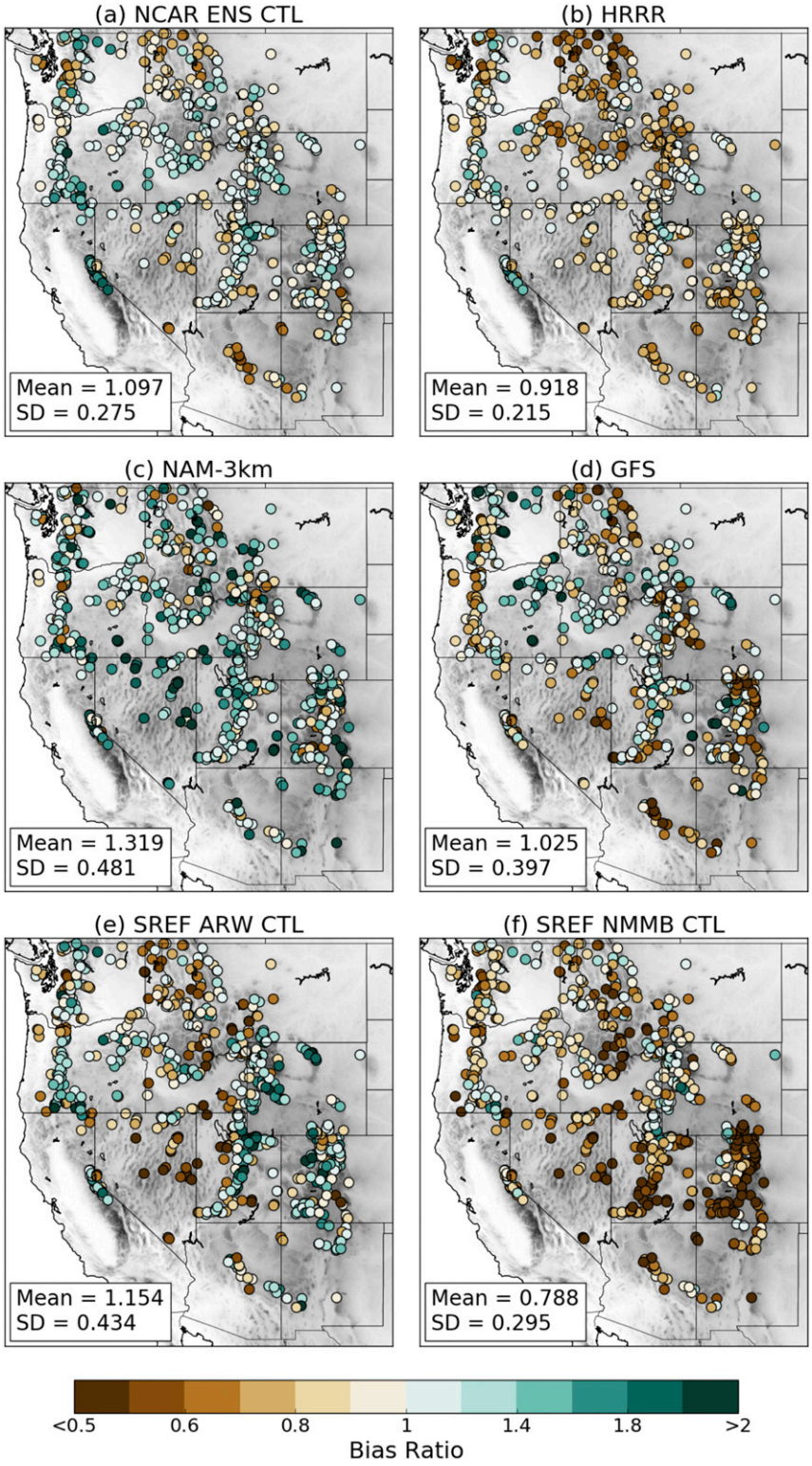

FIG. 5. Bias ratios at SNOTEL sites (color scale at bottom) and 30-arc-s topography (as in Fig. 2a) with mean bias ratio and standard deviation (SD) annotated: (a) NCAR ENS CTL, (b) HRRR, (c) NAM-3km, (d) GFS, (e) SREF ARW CTL, and (f) SREF NMMB CTL.

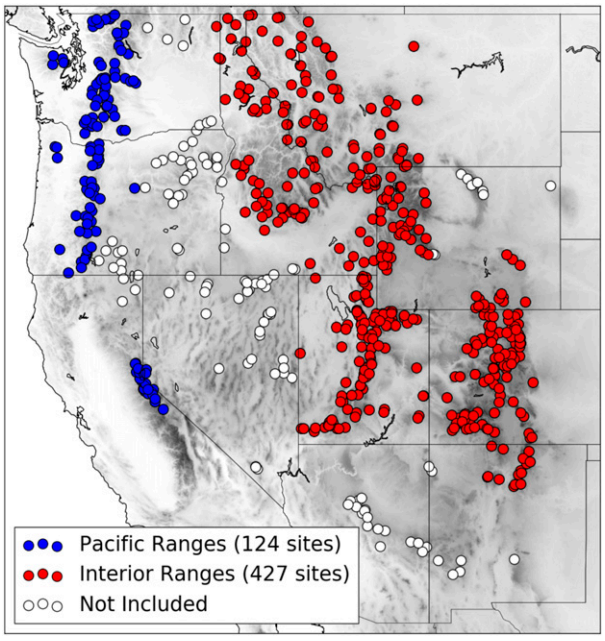

FIG. 6. Regional classification of SNOTEL sites and 30-arc-s topography (as in Fig. 2a).

observed in both regions, while the SREF NMMB CTL produced the least total precipitation in both regions, including only ~78% of total observed precipitation over the Pacific ranges. Overall, these results are consistent with Fig. 5.

Bias ratios computed relative to PRISM analyses illustrate some of the spatial characteristics of precipitation forecasts over the western United States. For brevity, we focus on bias ratios over the complex terrain surrounding Salt Lake City, Utah (SLC), and Lake Tahoe, California. In the region surrounding SLC, bias ratios produced by the NCAR ENS CTL, HRRR, and NAM-3km generally increase from west (windward side) to east (leeward side) across the Stansbury Mountains, Oquirrh Mountains, and Wasatch Range (Figs. 8a–c). The NCAR ENS and HRRR, for example, produce bias ratios < 1 on the western slopes, ~1 near the crests, and >1 on the eastern slopes of these mountain ranges (Figs. 8a,b). Although the NAM-3km has a wet bias over the eastern and western slopes of all three ranges, its local bias ratio maxima are on the eastern slopes, consistent with a bias ratio increase from west to east (Fig. 8c). Despite poorly resolving the three ranges, the GFS also exhibits a general tendency for the bias ratio to increase from the windward to leeward slopes (Fig. 8d). The SREF ARW CTL and SREF NMMB CTL overpredict valley precipitation and underpredict mountain precipitation (Figs. 8e,f).

In the region surrounding Lake Tahoe, bias ratios produced by the NCAR ENS CTL, HRRR, and NAM-3km similarly increase from west to east across the Sierra Crest, Carson Range, and Pine Nut Mountains, with all three models exhibiting pronounced wet biases on their eastern (leeward) slopes (Figs. 9a–c). Bias ratios for the GFS, SREF ARW CTL, and SREF NMMB CTL exhibit minimal topographic dependence over the Sierra Crest and are generally <1 over the Carson Range and Pine Nut Mountains (Figs. 9d–f).

Overall, the cross-barrier characteristics evident above broadly represent spatial bias ratio characteristics across the west. Although the mean bias ratio varies regionally, NCAR ENS, HRRR, and NAM-3km bias ratios typically increase as one moves climatologically downstream across mountain barriers, which could reflect either systematic biases in these modeling systems or PRISM analysis methods. If this reflects a model bias, it may be the result of terrain smoothing leading to poorly resolved orographic processes and/or deficiencies in microphysical parameterizations that allow too much precipitation to be carried over mountain crests. Aside from a dry bias over very narrow mountain ranges (i.e., Carson Range), spatial bias ratio characteristics in the lower-resolution GFS, SREF ARW CTL, and SREF NMMB CTL are less generalizable, likely because very narrow mountain ranges are not resolved and wider mountain ranges are inadequately represented.

Next, we bin events (2.54-mm intervals) to examine the ratio of forecast to SNOTEL-observed event frequencies (i.e., frequency bias) as a function of event size (Fig. 10). We assume frequency biases > 1.2 reflect a clear overprediction of event frequency and <0.85 a clear underprediction. Except for the NAM-3km, which overpredicts events > 36 mm, and SREF NMMB CTL, which underpredicts events < 30 mm, all models generally exhibit frequency biases between 0.85 and 1.2 for all event sizes in the Pacific ranges (Fig. 10a). Aside from the HRRR, frequency bias scores are generally worse over the interior ranges (Fig. 10b). The NCAR ENS CTL overpredicts events > 28 mm and the NAM-3km overpredicts events > 18 mm. The NAM-3km overprediction grows nearly monotonically with event size, with a frequency bias > 2 for events > 39 mm. The GFS exhibits better frequency biases than the NCAR ENS CTL and NAM-3km, but the GFS underpredicts events $>$ 42 mm. Except for an overprediction of events > 42 mm, the SREF ARW CTL generally displays no clear signs of overprediction or underprediction. The SREF NMMB CTL significantly underforecasts the frequency of events < 22 mm and overforecasts the frequency of events > 38 mm.

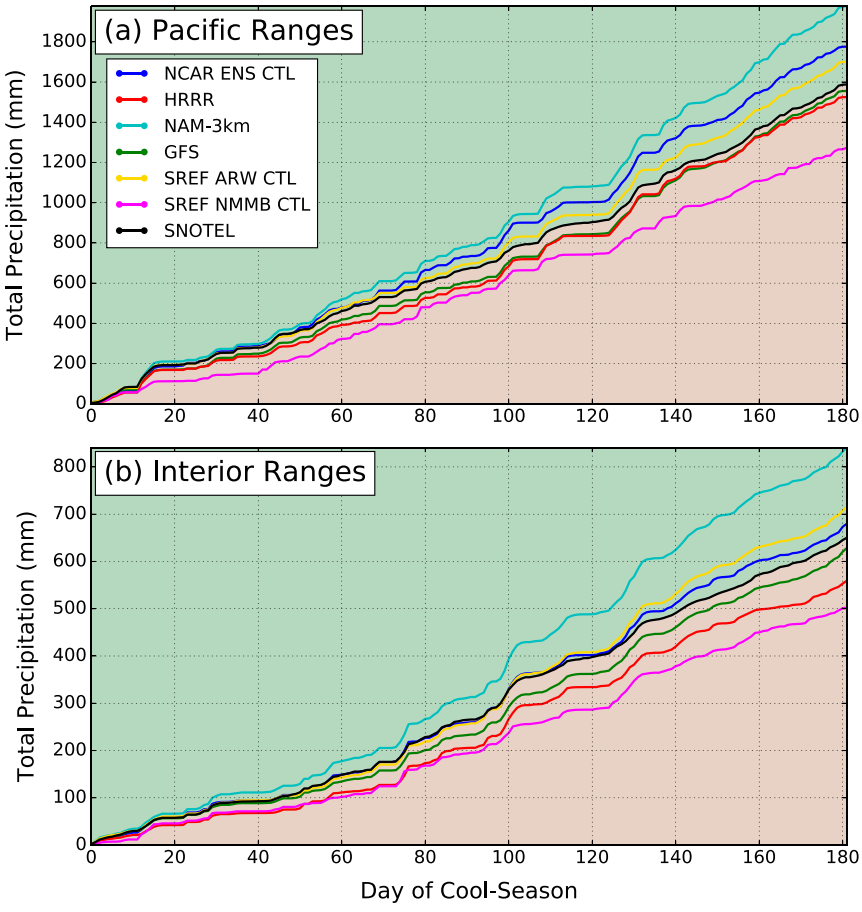

FIG. 7. Mean observed and forecast accumulated cool-season precipitation at SNOTEL sites in the (a) Pacific ranges and (b) interior ranges. Light green (light brown) shading indicates values above (below) the SNOTEL mean.

Overall, we find the least bias present in the NCAR ENS CTL and HRRR. Both models produce accurate cool-season precipitation totals at most SNOTEL sites. A slight wet bias in the NCAR ENS CTL and dry bias in the HRRR is revealed when looking at total precipitation averaged over both regions. Aside from the NCAR ENS CTL producing too many large events in the interior ranges, both models generate an accurate number of events. Conversely, the NAM-3km exhibits a significant wet bias at most SNOTEL sites, while the GFS and SREF ARW CTL have minimal bias for cumulative SNOTEL site statistics, but a substantial wet or dry bias from site to site. The site-to-site variations in bias may at least partially reflect terrain smoothing, as discussed above. A dry bias due to too few small and moderate events is found in the SREF NMMB CTL.

3) DISTRIBUTIONS OF FORECAST EVENTS

We now focus on forecasts and their corresponding observations (i.e., event pairs) using bivariate histograms (Fig. 11). Bias is reflected by frequent event pairs falling above (underprediciton) or below (overprediction) the 1:1 line, while precision is reflected by limited scatter of event pairs. Ideally, a model has minimal bias and high precision. Median values for each dimension of the bivariate histogram are plotted to help interpretation. In both regions, the NCAR ENS CTL has frequent event pairs falling near the 1:1 line, indicating minimal bias and moderate precision (Figs. 11a,g). Aside from a slight tendency for event pairs < 20 mm to fall above the 1:1 line (underprediciton) in the Pacific ranges, the HRRR displays minimal bias and greater precision than the NCAR ENS CTL (Figs. 11b,h). Consistent with its previously discussed wet bias, the NAM-3km has frequent event pairs falling below the 1:1 line for all event sizes in both regions, reflecting frequent overprediction (Fig. 11c,i). The GFS exhibits minimal bias and moderate precision, similar to the NCAR ENS CTL (Figs. 11d,j). The SREF ARW CTL displays low precision in both regions and overprediction for events < 22 mm (Figs. 11e,k). Low

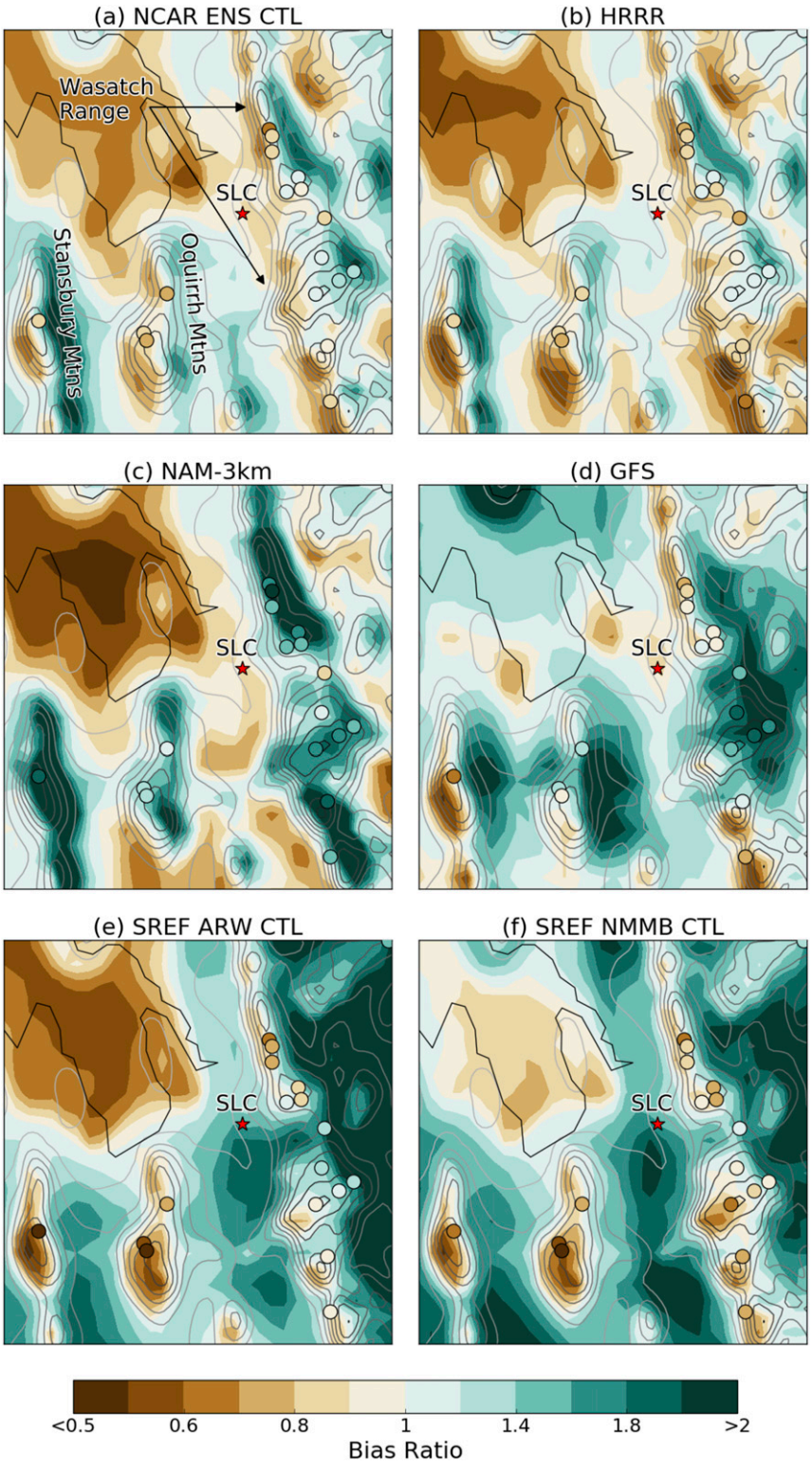

FIG. 8. Bias ratios relative to PRISM analyses (following scale at bottom) and SNOTEL observations (filled circles following scale at bottom) in the region surrounding SLC for the (a) NCAR ENS CTL, (b) HRRR, (c) NAM-3km, (d) GFS, (e) SREF ARW CTL, and (f) SREF NMMB CTL. The 1-arc-min topography is smoothed using a rectangular smoother and contoured every 200 m from 1300 (light gray) to 3300 m MSL (black). Mountain ranges are annotated in (a).

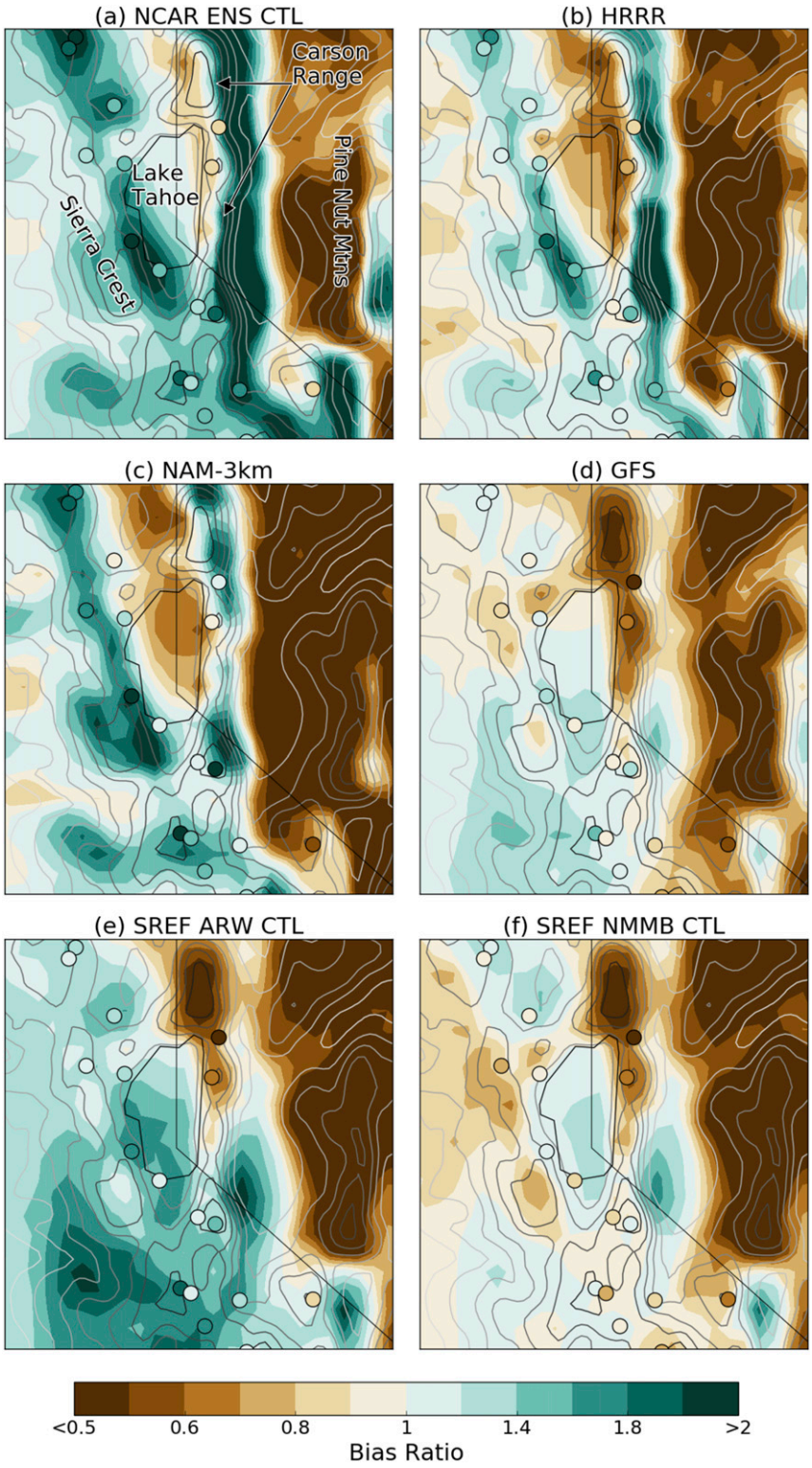

FIG. 9. As in Fig. 8, but for the Lake Tahoe region and topography contoured every 200 m from 1000 (light gray) to 2800 m MSL (black). Lake Tahoe is annotated for reference in (a).

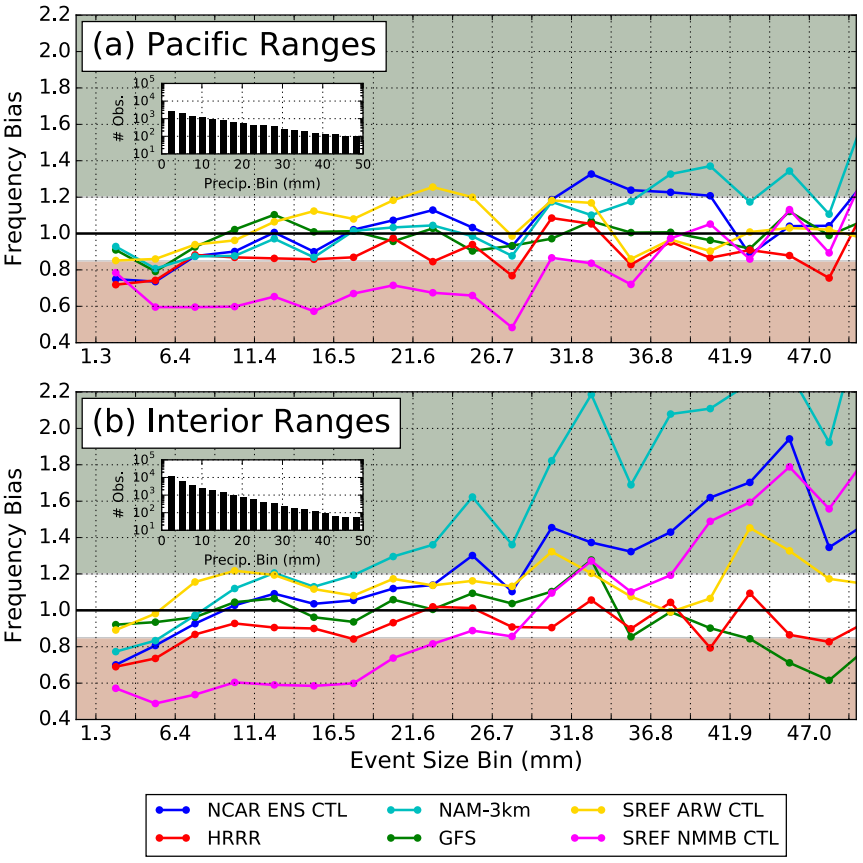

FIG. 10. Frequency bias as a function of event size at SNOTEL sites in the (a) Pacific ranges and (b) interior ranges. Green (brown) shading indicates bias ratios ≥ 1.2 (≤0.85). Samples size in each bin is shown in inset histograms.

precision is shown by the SREF NMMB CTL with minimal correlation between forecasts and observations (Figs. 11f,l). Considering that high precision and minimal bias indicate good accuracy, the HRRR features the greatest accuracy, followed by the NCAR ENS CTL and GFS. The NAM 3-km frequently overpredicts events, leading to poor accuracy, while the SREF ARW CTL and especially the SREF NMMB CTL are characterized by low precision and are least accurate.

b. *Deterministic accuracy measures*

We now evaluate statistical measures based on a standard 2 × 2 contingency table using absolute event thresholds to determine model performance characteristics as a function of event size. Aided by its wet bias, the NAM-3km scores the highest hit rates over both regions for all event thresholds (>0.6 over Pacific ranges and generally >0.4 over interior ranges; Figs. 12a,b). The NCAR ENS CTL, HRRR, GFS, and SREF ARW CTL produce similar hit rates for event thresholds < 23 mm, while the NCAR ENS CTL and HRRR score slightly higher than the GFS and SREF ARW CTL for event thresholds > 23 mm over the Pacific ranges. Over the interior ranges, the NCAR ENS CTL's hit rate improves relative to other models and is greater than or equal to the HRRR's for all event thresholds (Fig. 12b). The hit rates for the GFS and SREF ARW CTL drop off considerably for event thresholds > 23 mm over the interior ranges. The SREF NMMB CTL performs poorly in both regions, recording hit rates < 0.5 for all event thresholds (Figs. 12a,b).

The HRRR produces the lowest false alarm ratios for all thresholds in both the Pacific and interior ranges (Figs. 12c,d). Again, we find a substantial improvement in the NCAR ENS CTL's scores over the interior ranges compared to the Pacific ranges (Figs. 12c,d); its false alarm ratio is relatively poor (>0.4 for event thresholds > 20 mm) and similar to that of the NAM-3km and SREF ARW CTL over the Pacific ranges, but improves relative to all other models and is similar to that of the GFS over the interior ranges (Figs. 12c,d). Even with its significant dry bias, the SREF NMMB CTL records the worst false

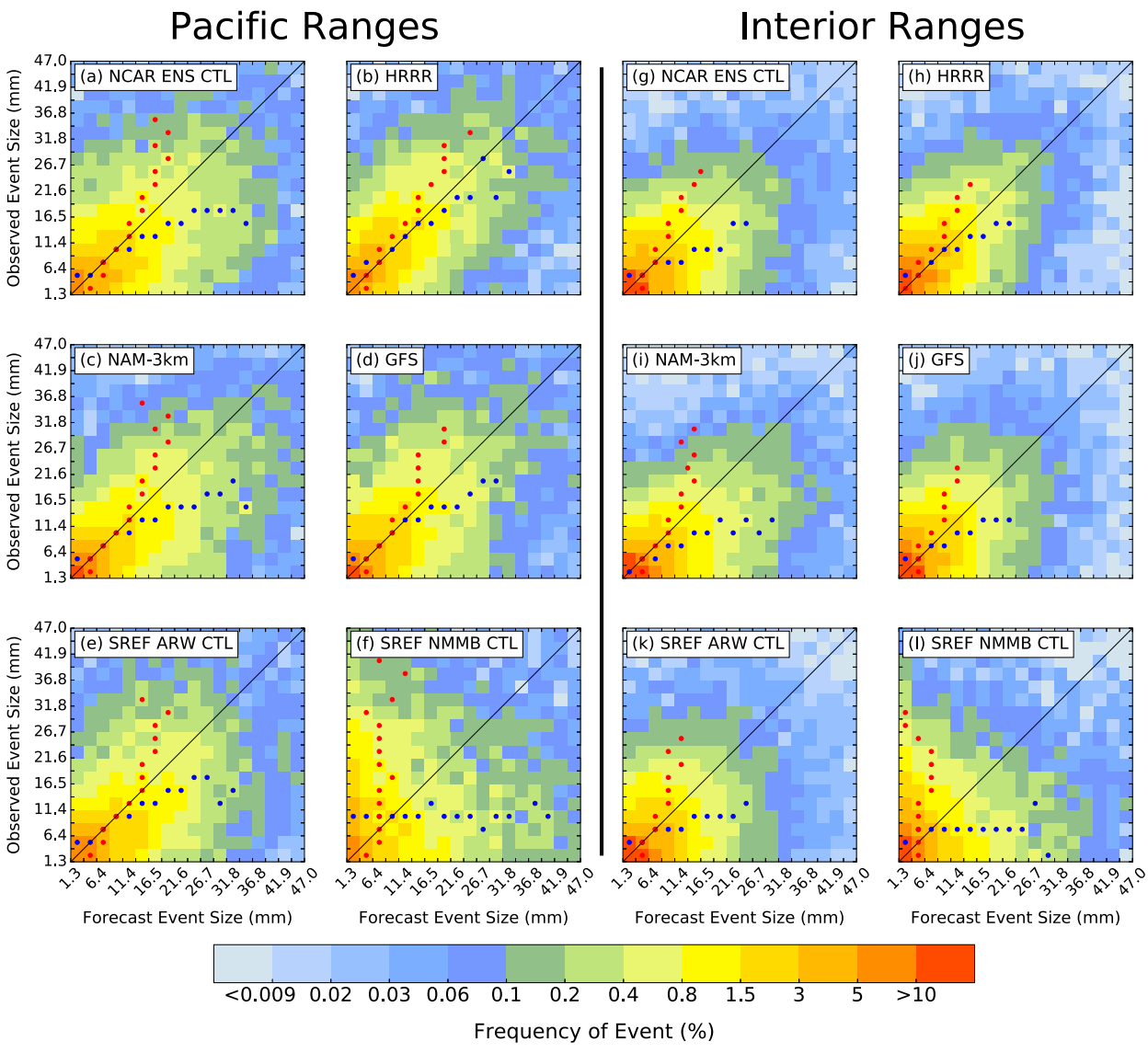

FIG. 11. Bivariate histograms of forecast and observed precipitation at SNOTEL sites in the Pacific ranges for the (a) NCAR ENS CTL, (b) HRRR, (c) NAM-3km, (d) GFS, (e) SREF ARW CTL, and (f) SREF NMMB CTL. (g)–(l) As in (a)–(f), but over the interior ranges. Red (blue) dots represent the median observed (forecast) event size in each bin. Dots are not shown for bins with fewer than 50 events.

alarm ratios for all event thresholds over both regions (Figs. 12c,d).

Over both the Pacific and interior ranges, the HRRR and NAM-3km generally produce the highest ETSs (Figs. 12e,f). Because models with larger biases tend to have higher ETSs (Mason 1989), the NAM-3km's ETS is likely aided by its wet bias. The GFS is more skillful (larger ETSs) than the NCAR ENS CTL over the Pacific ranges but is less skillful (smaller ETSs) over the interior ranges. Consistent with other statistical measures, the SREF ARW CTL and especially the SREF NMMB CTL exhibit less skill over both ranges (Figs. 12e,f). A general decline in ETSs by all models is evident over the interior ranges, especially for event thresholds > 25 mm. Overall, the highest-resolution deterministic models perform best, as they are able to better resolve the terrain and thus orographic precipitation. The NCAR ENS CTL may have less skill relative to all other models over the Pacific ranges compared to the interior ranges because the western boundary of its 3-km forecast domain is very close to the Pacific coast relative to the other models (Fig. 1a).

We now focus on the same deterministic measures using upper-quartile and greater percentile event thresholds to evaluate bias-corrected model performance. Percentiles computed from SNOTEL observations and model forecasts reveal biases consistent with previous results (Fig. 13).

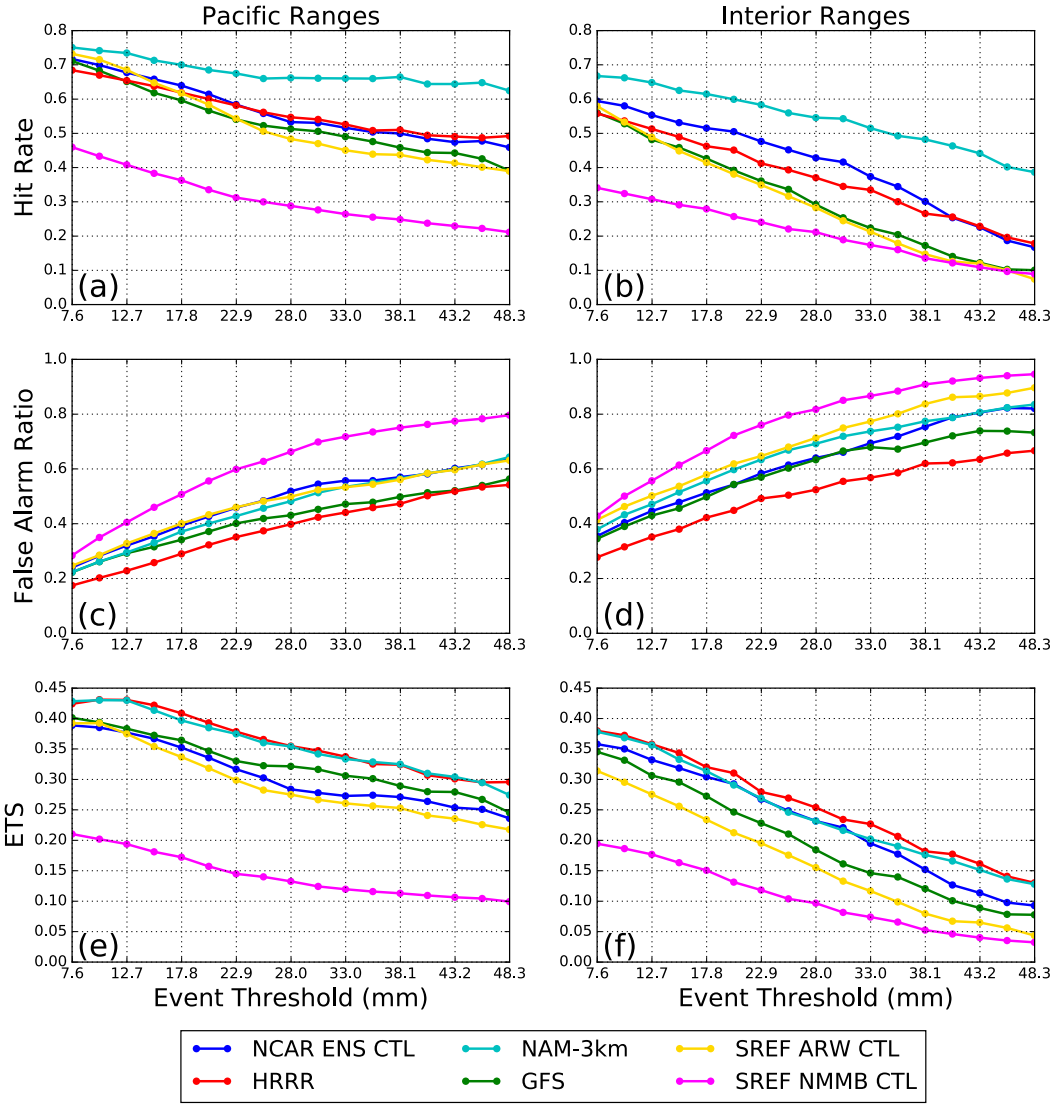

FIG. 12. Verification metrics based on Table 2 as a function of absolute event thresholds (mm) at SNOTEL sites. (a) Hit rate in the Pacific ranges. (b) Hit rate in the interior ranges. (c),(d) As in (a),(b), but for false alarm ratio. (e),(f) As in (a),(b), but for ETS.

In general, bias correction improves the hit rate of models with a dry bias (i.e., the HRRR) and reduces the hit rate of models with a wet bias (i.e., the NAM-3km). Therefore, the HRRR exhibits the highest hit rates in both regions, followed by the NAM-3km and GFS in the Pacific ranges and the NAM-3km and NCAR ENS CTL in the interior ranges (Figs. 14a,b). Contrary to the effect of bias correction on hit rates, false alarm ratios worsen (increase) for models with a dry bias and improve (decrease) for models with a wet bias when bias correction is applied (Figs. 14c,d). The impact of removing bias on ETS is subtler, but we do find slight improvements in the scores of models with a dry bias and slight declines in the scores of models with a wet bias, such that the HRRR produces higher ETSs than the NAM-3km over both regions for almost all thresholds (Figs. 14e,f). The relative decrease in all three metrics for all models at the 85th percentile threshold over the interior ranges (Figs. 14b,d,f) is due to the discrete nature of SNOTEL data, which results in the same absolute threshold corresponding to a range of percentile thresholds (Fig. 13b). This does not affect intraregional model comparisons. Overall, we find the bias-corrected results (Fig. 14) to be generally consistent with the non-bias-corrected results (Fig. 12) when accounting for the impact that bias has on these three statistical measures.

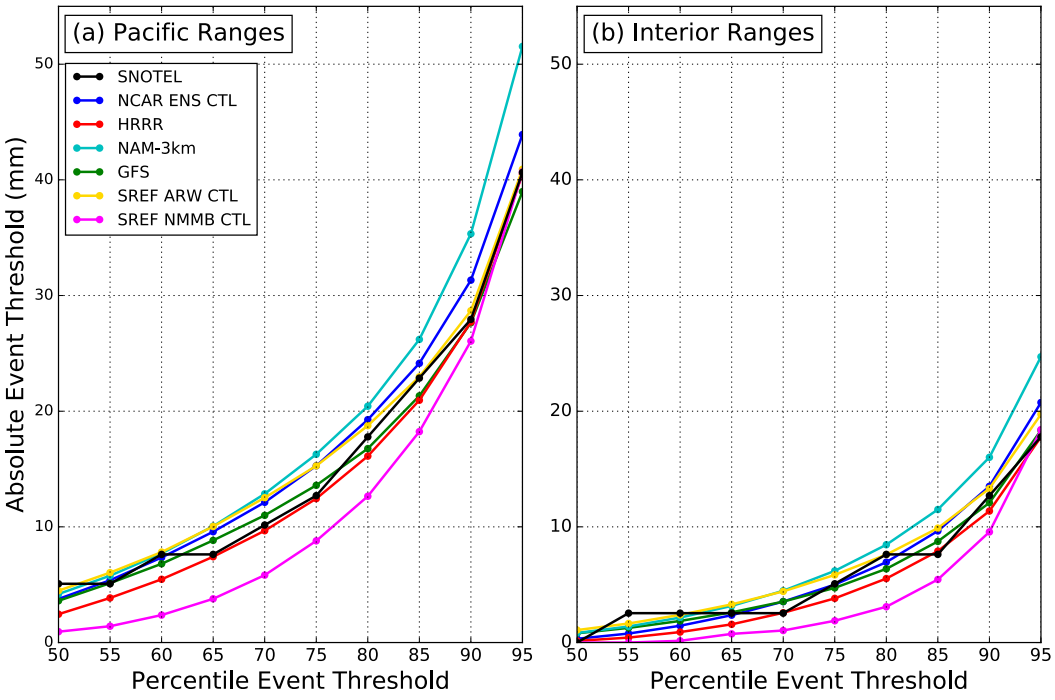

FIG. 13. Forecast and observed absolute event thresholds (mm) corresponding to percentile thresholds for all forecast and observed events at SNOTEL sites in the (a) Pacific ranges and (b) interior ranges.

The EM and PMM can be calculated to simplify ensemble information into a single forecast (Clark 2017). An evaluation of the performance of the NCAR ENS, SREF ARW, and SREF NMMB CTL, EM, and PMM using the same bias-corrected, deterministic measures as above reveals general improvement relative to their respective control members (Fig. 15). The NCAR ENS and SREF NMMB EM and PMM produced higher hit rates, lower false alarm ratios, and higher ETSs than their control members for all percentile thresholds in both regions. Although less discernible and not true for all percentile thresholds, this behavior is largely the case for the SREF ARW CTL, EM, and PMM as well. Differences between the EM and PMM for all ensembles are negligible, possibly because the utility of the PMM in restoring amplitude to the EM precipitation field is reduced when the precipitation forcing mechanism is static (i.e., terrain). The improvement in the SREF NMMB EM and PMM over its control is so large that they exhibit the best ETSs for all event percentile thresholds over the Pacific ranges (Fig. 15e). This improvement in the SREF NMMB and relative lack of improvement in the NCAR ENS and SREF ARW is likely due to significant spread in the SREF NMMB and minimal spread in the NCAR ENS and SREF ARW, as shown in section 3c.

c. *Probabilistic verification*

Similar to the method used for bias-corrected, deterministic validation, we examine PQPFs from the NCAR ENS and SREF using percentile event thresholds. Non-bias-corrected PQPFs were also inspected and produced very similar results (not shown). Ideally, each member of an ensemble should be equally likely to be closest to the "truth," and, thus, all members should have similar climatologies. A tight packing of precipitation distributions for each member of the NCAR ENS reveals that all members indeed have similar climatologies, confirming the expectation of equal likelihood, where the climatologies are characterized by a wet bias for the 80th percentile and larger events in both regions (Fig. 16). Conversely, an exceptional bifurcation is present in the distributions of SREF members as a result of its use of two dynamical cores. Clearly, the design of the SREF violates the principal of equal likelihood. Members within each core also exhibit greater spread than the NCAR ENS, reflecting the use of multiple physical parameterizations in the SREF (Table 1). While the SREF ARW members contain a wet bias, the SREF NMMB members exhibit a sizeable dry bias, especially for 85th percentile events and smaller (Fig. 16). Because of the dramatic differences in the climatologies of the two SREF cores, we examine the performance of the individual cores in addition to the entire 26-member SREF.

We use attributes diagrams to assess the probabilistic performance of the NCAR ENS, SREF, and the individual dynamical cores of the SREF (SREF ARW and

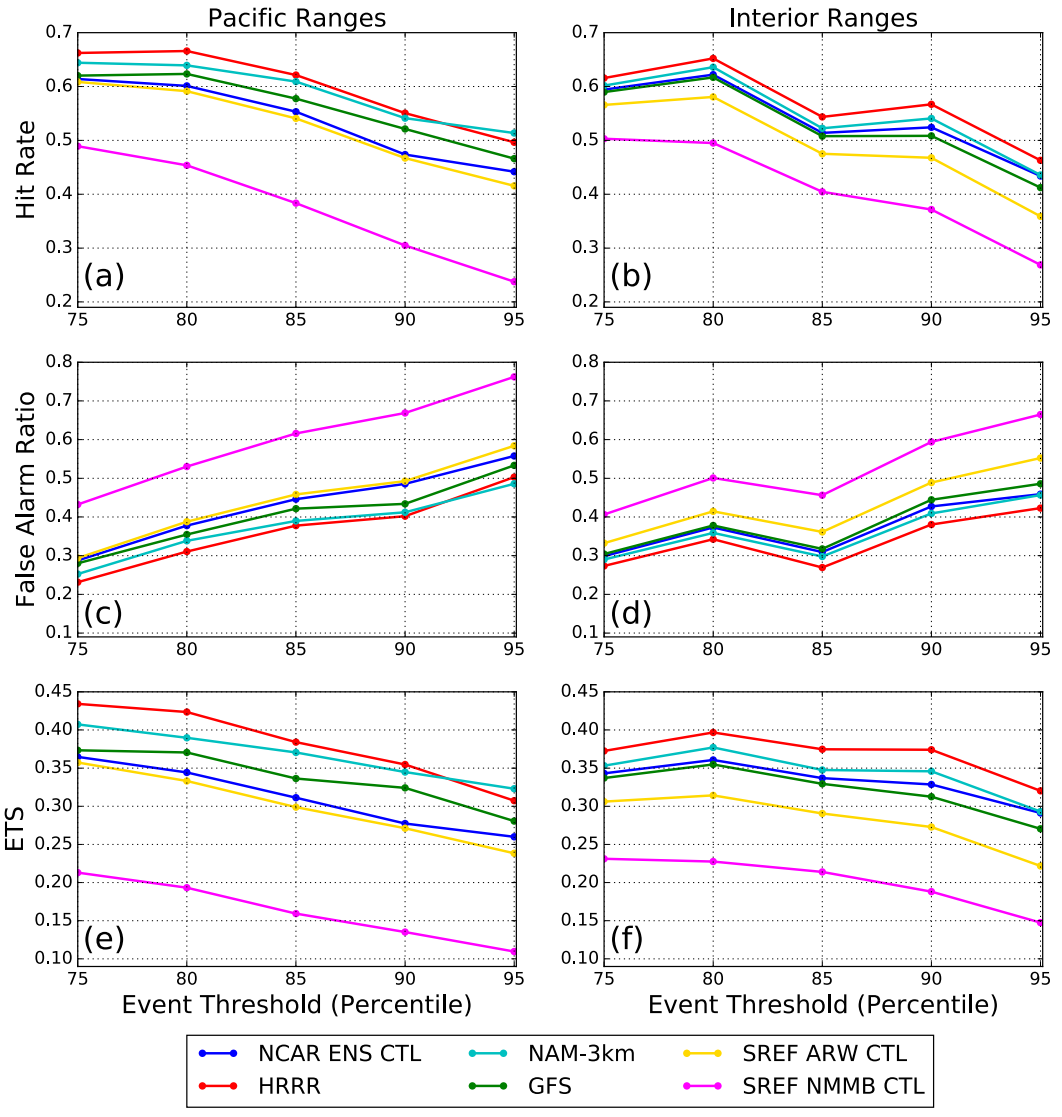

FIG. 14. As in Fig. 12, but based on percentile thresholds (Fig. 13).

the SREF NMMB) at forecasting 85th and 95th percentile events. Attributes diagrams provide information regarding the Brier score decomposition (reliability, resolution, and uncertainty) and other characteristics of each ensemble. Forecast probability bins of 0%–5%, 5%–15%, . . . , 85%–95%, and 95%–100% are used to construct and compute attributes diagrams and statistics. The shapes of the reliability curves for the SREF and especially the NCAR ENS for the 85th percentile events in both regions display overconfidence (Figs. 17a,b). For example, over the Pacific ranges, when the NCAR ENS forecasts a 90% probability that an 85th percentile event will occur, it only occurs ~67% of the time, whereas when it forecasts a 10% probability that the event will occur, it occurs ~24% of the time (Fig. 17a). The SREF has better reliability in both regions and better resolution over the Pacific ranges, leading to higher BSSs (0.340 57 over the Pacific ranges and 0.316 98 over the interior ranges) than the NCAR ENS (0.296 01 over the Pacific ranges and 0.313 83 over the interior ranges; Table 3). Although the BSS equally weighs reliability and resolution, resolution is considered the most important attribute of an ensemble (Toth et al. 2003). While reliability can be improved using a posteriori calibration techniques, resolution cannot and can only be increased by a clearer segregation of scenarios where the event of interest occurs with higher or lower frequency than climatology (i.e., a better forecast in a probabilistic sense). The forecast frequency histograms reveal that the NCAR ENS forecasts high or low

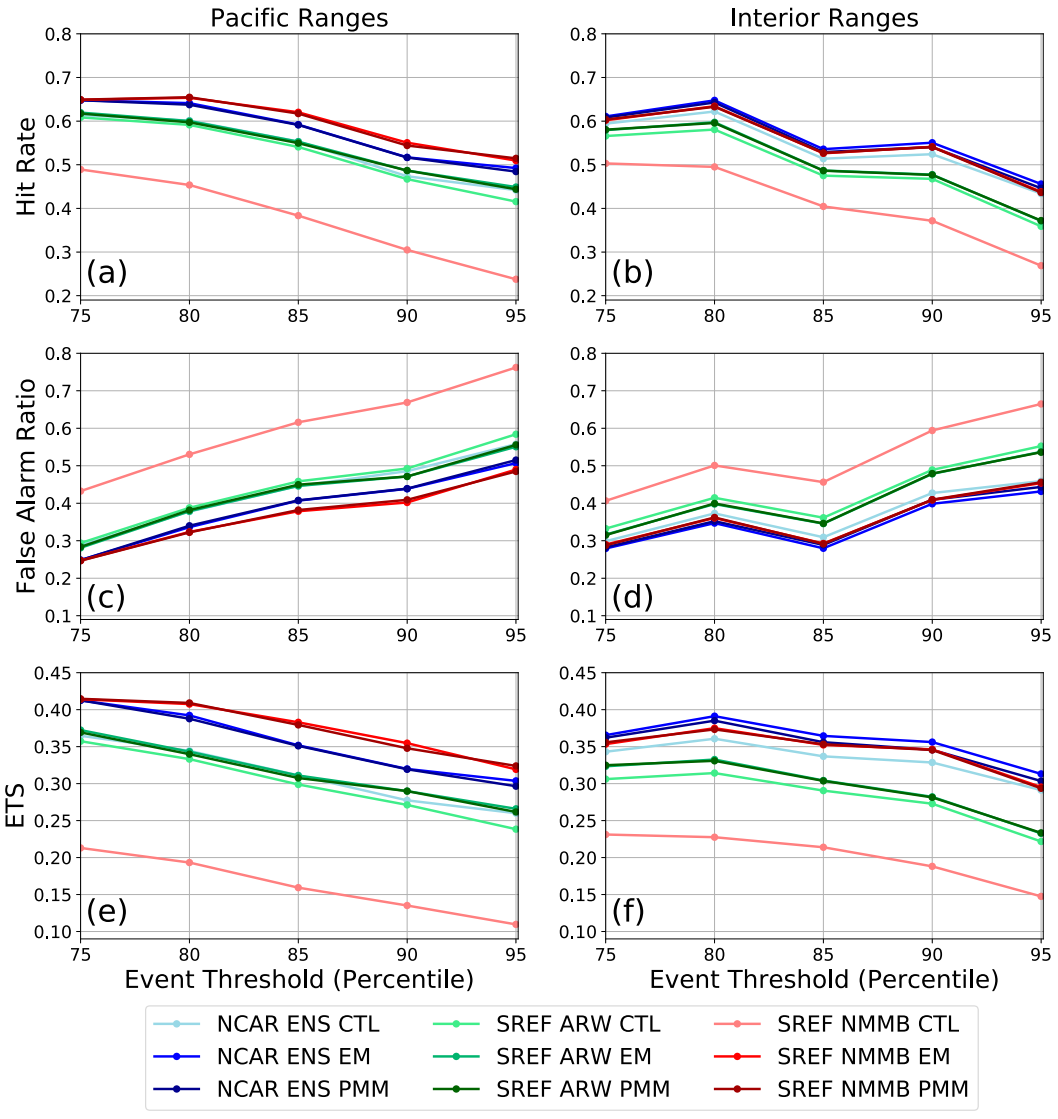

FIG. 15. As in Fig. 14, but for NCAR ENS, SREF ARW, and SREF NMMB CTL, EM, and PMM.

probabilities more often than the SREF, indicating better sharpness, the tendency of an EPS to produce forecasts near 0 or 1 (Murphy 1993; Figs. 17a,b). However, this better sharpness is accompanied by overconfidence and relatively poor reliability, indicating that the NCAR ENS is likely spread deficient.

We find similar performance characteristics in the NCAR ENS and SREF when focusing on the 95th percentile event thresholds (Figs. 18a,b and Table 4). Overconfidence is again evident in both ensembles, although to a lesser extent than at the 85th percentile threshold. While the SREF continues to outperform the NCAR ENS over the Pacific ranges, with better reliability and resolution (Fig. 18a and Table 4), the NCAR ENS produces a larger BSS over the interior, aided by good resolution (Fig. 18b and Table 4). The NCAR ENS forecasts probabilities of 1 more than twice as often as the SREF over the interior ranges, indicating more sharpness (Fig. 18b).

Although the SREF has a much coarser horizontal grid spacing (16 km) than the NCAR ENS (3 km), bias-corrected PQPFs from the full 26-member SREF are often more skillful, especially over the Pacific ranges. While the NCAR ENS is relatively sharp, it is unreliable as a result of insufficient spread. Conversely, the SREF contains more spread, which arises largely because of its two climatologically contrasting dynamical cores. Thus, even though this enhanced spread improves SREF verification metrics and suggests it is possible to construct a convection-parameterizing EPS that verifies

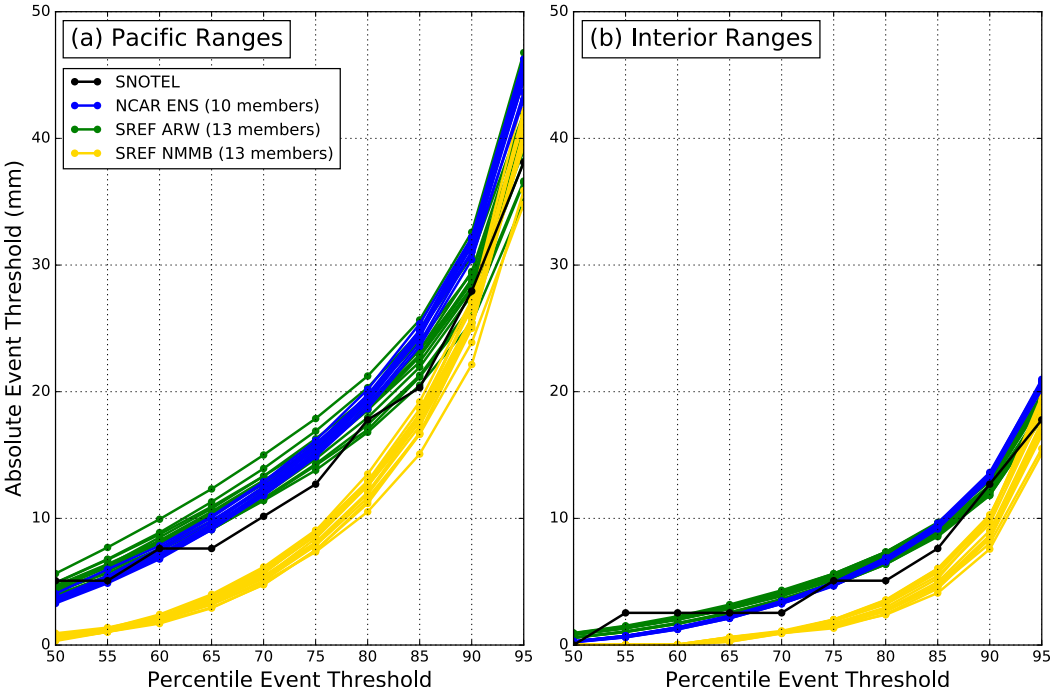

FIG. 16. As in Fig. 13, but for all members of the NCAR ENS, SREF ARW, and SREF NMMB.

better than a CPE over mountainous terrain, the SREF design clearly violates the principal of equal likelihood, which muddles the interpretation of its PQPFs. For example, it is inaccurate to state that there is a 70% chance of an event occurring when 70% of SREF members forecast the event to occur. However, interpretation of PQPFs from equally likely ensembles (like the NCAR ENS) is straightforward, but further developments are needed to improve spread in these types of EPSs (e.g., Romine et al. 2014).

Evaluating the performance and characteristics of the two 13-member SREF cores (SREF ARW and SREF NMMB) provides insights into the characteristics of the full, 26-member SREF. Under all scenarios (85th and 95th percentile event thresholds in both regions), the SREF NMMB exhibits better reliability and resolution and, hence, a larger BSS, than the SREF ARW (Figs. 17c,d and 18c,d and Tables 3 and 4). The SREF ARW suffers from significant overconfidence under all scenarios. Frequency histograms reveal a lack of sharpness (large spread) in the SREF NMMB, especially over the interior ranges for 85th and 95th percentile event thresholds (Figs. 17d and 18d). Given that one would not expect an individual member of an ensemble with large spread to perform well deterministically, the large SREF NMMB spread corresponds well with the poor performance of the SREF NMMB CTL.

Although the 26-member SREF is generally more skillful than the NCAR ENS, the NCAR ENS often outperforms the individual SREF dynamical cores; the NCAR ENS is more skillful than the SREF ARW over the entire western United States and the SREF NMMB over the interior ranges (Tables 3 and 4; see BSSs). Therefore, when examining the 13-member SREF subensembles, the advantage of a CPE compared to a convection-parameterizing EPS becomes more apparent, which indicates that the full 26-member SREF likely outperformed the NCAR ENS primarily because forecasts from two dynamical cores were combined.

4. Conclusions

This study has evaluated the performance of precipitation forecasts from the convection-permitting NCAR ENS and several operational forecast systems at high-elevation SNOTEL sites across the western United States during the 2016/17 cool season. The NCAR ENS CTL and HRRR exhibit superior precipitation biases as evinced by the ratio of forecast-to-observed mean daily precipitation and the ratio of forecast-to-observed event frequencies. Because the HRRR precipitation forecasts are effectively a combination of two short-term forecasts, the HRRR may have had an advantage compared to the other NWP systems. The GFS and SREF ARW CTL produce minimal overall bias but overpredicted or underpredicted

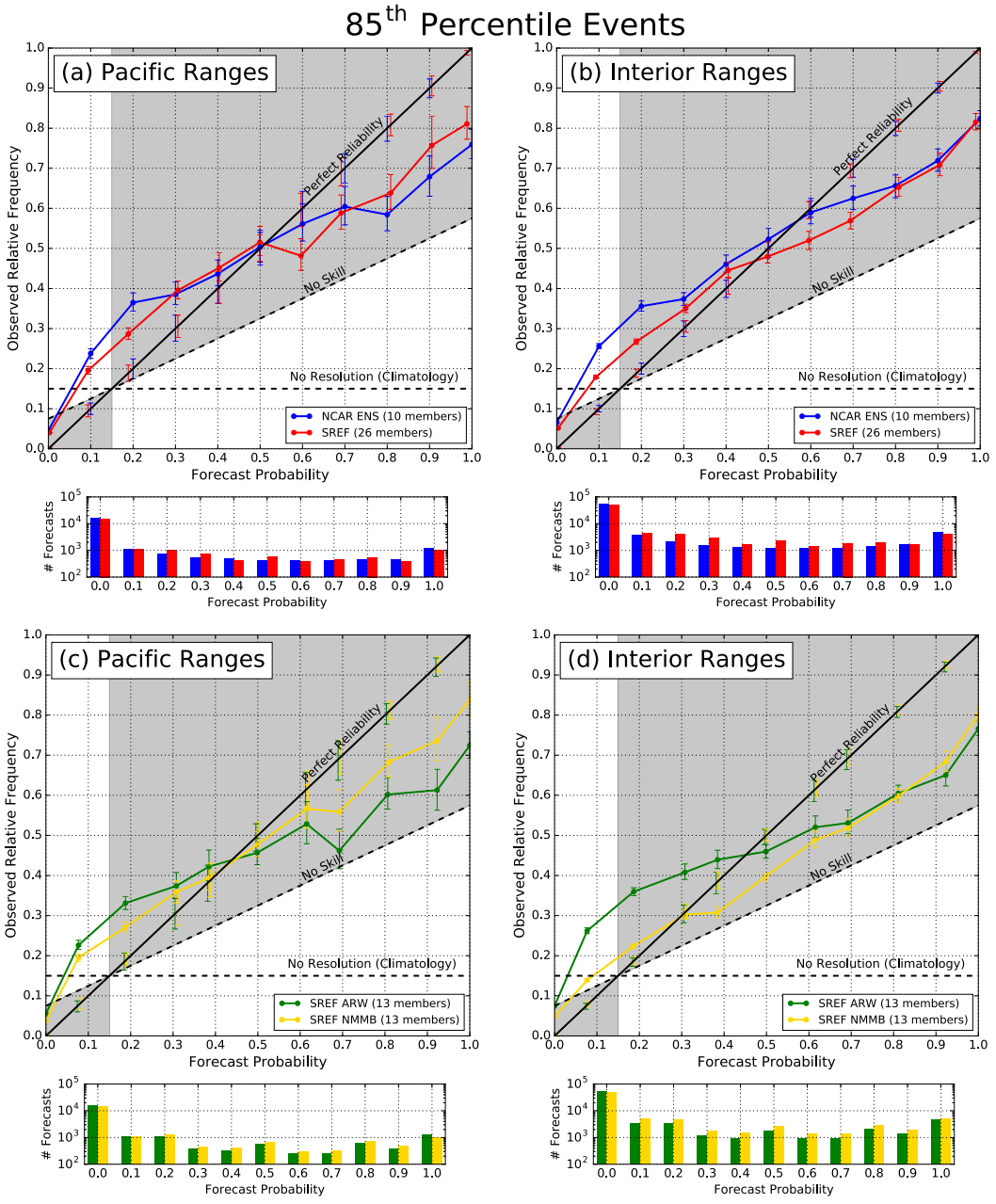

FIG. 17. Attributes diagrams for NCAR ENS and SREF forecast and SNOTEL observed 85th percentile events in the (a) Pacific ranges and (b) interior ranges. Gray shadings indicate those probability bins that contribute positively to the BSS with a reference of climatology. The 95% consistency bars and confidence intervals are shown on the perfect reliability line and plotted reliability line, respectively. (c),(d) As in (a),(b), but for SREF ARW and SREF NMMB. Forecast frequency histograms at bottom indicate number of forecasts in each forecast probability bin.

precipitation on a site by site basis. A significant wet bias is revealed in the NAM-3km due to its tendency to produce too many large events, especially over the interior ranges for events $\geq$ 20 mm, whereas the SREF NMMB CTL generates too few moderate and small events $\leq$ 20 mm over both regions, yielding a substantial dry bias.

Deterministic validation metrics (i.e., equitable threat scores, hit rates, and false alarm ratios) using absolute event thresholds indicate that the higher-resolution NCAR ENS CTL, HRRR, and NAM-3km generally perform better than the coarser GFS, SREF ARW CTL, and SREF NMMB CTL. One exception is the performance of the

TABLE 3. Probabilistic metrics for 85th percentile events.

		BS	Reliability	Resolution	Uncertainty	BSS
Pacific ranges	NCAR ENS	0.101 05	0.008 81	0.051 30	0.143 54	0.296 01
	SREF	0.094 66	0.005 26	0.054 15	0.143 54	0.340 57
	SREF ARW	0.112 01	0.012 45	0.043 98	0.143 54	0.219 66
	SREF NMMB	0.091 15	0.004 95	0.057 34	0.143 54	0.364 99
Interior ranges	NCAR ENS	0.107 15	0.008 37	0.057 37	0.156 15	0.313 83
	SREF	0.106 66	0.005 96	0.055 46	0.156 15	0.316 98
	SREF ARW	0.123 60	0.013 52	0.046 08	0.156 15	0.208 49
	SREF NMMB	0.108 95	0.009 00	0.056 20	0.156 15	0.302 26

NCAR ENS CTL over the Pacific ranges, where it exhibits inferior ETSs and false alarm ratios than the GFS. The SREF ARW CTL generally performs second worst for all three metrics, while SREF NMMB CTL produces the worst scores by a significant margin for all three metrics in both regions. Consistent with other studies (e.g., Lewis et al. 2017), the performance of all six models declines from the Pacific to interior ranges.

We further bias correct these deterministic validation metrics by using percentile event thresholds. The removal of bias allows for a robust assessment of the placement of precipitation within the context of each model's climatology. Overall, the bias-corrected results are generally consistent with the non-bias-corrected results when accounting for the impact that bias has on the deterministic validation measures. For example, although the bias-corrected ETSs are slightly lower for models with a wet bias (i.e., the NAM-3km), we still find the HRRR, NAM-3km, and GFS to exhibit the highest ETSs over the Pacific ranges and the HRRR, NAM-3km, and NCAR ENS CTL to exhibit the highest ETSs over the interior ranges.

Prior studies note varied results concerning the benefits of decreasing horizontal grid spacing below 12 km over the western United States (Mass et al. 2002; Grubišić et al. 2005; Hart et al. 2005). Our results indicate that decreasing horizontal grid spacing to 3 km increases the performance of cool-season QPFs, especially over the interior ranges of the western United States. The importance of increased resolution over the interior ranges may reflect their narrow nature, whereas the Pacific ranges have a more sustained high-mountain mass and are better resolved at coarser resolutions. Additionally, precipitation systems over the Pacific ranges are more spatially coherent (Serreze et al. 2001), which may also enhance predictability.

Deterministic validation of the NCAR ENS, SREF NMMB, and SREF ARW EM and PMM show improvement in the EM and PMM over each ensemble's control member. While past studies focused on areas of flat terrain found the EM to poorly predict precipitation because it dampens high-amplitude features (e.g., Ebert 2001; Schwartz et al. 2014), our findings suggest that this is not true over complex terrain. This is likely because high-amplitude precipitation features in this study are primarily forced by terrain, which is represented similarly by all ensemble members. We find a lack of improvement in the PMM over the EM, which conflicts with the findings of several studies (e.g., Clark et al. 2009; Schwartz et al. 2014) and may be because the amplitude of the orographic precipitation is captured relatively well by the EM, as discussed above.

Although the NCAR ENS and SREF are both designed to produce short-range probabilistic forecasts, their configurations, characteristics, and biases are drastically different. While the NCAR ENS contains a single dynamical core and each member has identical physics, the SREF contains two dynamical cores (SREF ARW and SREF NMMB) with varied physics among the members in each core. Ideally, each member of an ensemble should be equally likely to be closest to the "truth," and, thus, all members should have similar climatologies. We find the precipitation climatologies for each member of the NCAR ENS to be similar, whereas the precipitation climatologies for the SREF bifurcate into two distinct clusters based on the dynamical core. Thus, while the NCAR ENS confirms the expectation of equal likelihood, the design of the SREF clearly violates this principal. Consistent with the biases of their control members, NCAR ENS members contain a slight wet bias for 80th percentile and larger events, SREF ARW members contain an overall slight wet bias, and SREF NMMB members exhibit a significant dry bias, especially for 85th percentile events and smaller.

Bias-corrected probabilistic validation metrics reveal the NCAR ENS is less skillful than the 26-member SREF over the Pacific ranges for both 85th and 95th percentile event thresholds and over the interior ranges for 85th percentile event thresholds. Meanwhile, the NCAR ENS exhibits more skill over the interior ranges for 95th percentile event thresholds compared to the 26-member

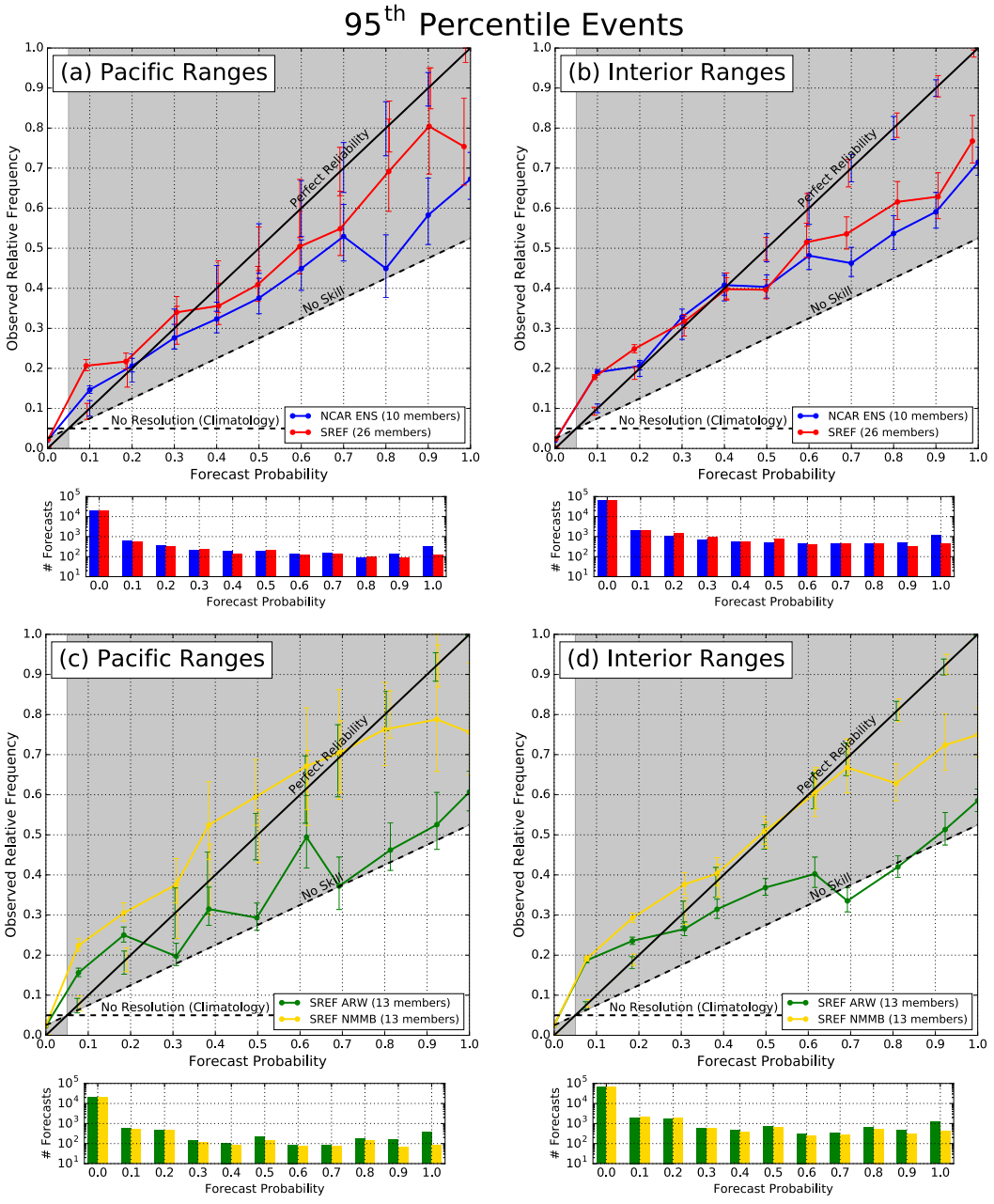

FIG. 18. As in Fig. 17, but for 95th percentile events.

SREF. Although probabilistic forecasts from the NCAR ENS are characterized by good sharpness, the NCAR ENS is overconfident and has poor reliability, whereas the 26-member SREF is less sharp and more reliable. Compared to bias-corrected PQPFs from 13-member subensembles composed of SREF members with the same dynamical core (SREF ARW and SREF NMMB), the NCAR ENS is more skillful than the SREF ARW over the entire western United States and the SREF NMMB over the interior ranges. Only by combining two ensemble systems with drastically different climatologies can the full 26-member SREF generate PQPFs that are generally more skillful than the NCAR ENS, especially over the Pacific ranges.

These findings indicate the advantages of high-resolution deterministic models and future promise of CPEs over the western United States. The HRRR, NAM-3km, and NCAR ENS CTL consistently outperform the coarser GFS, SREF ARW CTL, and SREF NMMB, especially over the interior ranges. As

TABLE 4. Probabilistic metrics for 95th percentile events.

		BS	Reliability	Resolution	Uncertainty	BSS
Pacific ranges	NCAR ENS	0.043 27	0.003 74	0.014 43	0.053 96	0.198 15
	SREF	0.040 77	0.001 48	0.014 67	0.053 96	0.244 42
	SREF ARW	0.048 14	0.006 48	0.012 30	0.053 96	0.107 88
	SREF NMMB	0.040 47	0.001 88	0.015 37	0.053 96	0.250 09
Interior ranges	NCAR ENS	0.042 93	0.003 48	0.016 84	0.056 29	0.237 38
	SREF	0.044 16	0.001 88	0.014 00	0.056 29	0.215 46
	SREF ARW	0.051 89	0.007 02	0.011 42	0.056 29	0.078 18
	SREF NMMB	0.043 58	0.001 99	0.014 70	0.056 29	0.225 75

computational resources increase, future work should focus on the development of operational deterministic models with horizontal grid spacings of 3 km or smaller. Although the NCAR ENS suffers from spread deficiency, its configuration could serve as a framework for the future development of short-range ensembles. With a horizontal grid spacing of 3 km, an individual member of the NCAR ENS is shown to be more skillful than two individual members of the 16-km SREF and, because the NCAR ENS follows the principal of equal likelihood, its probabilistic forecasts can be easily interpreted. The NCAR ENS's shortcoming is insufficient spread, which, although common in CPEs (e.g., Clark et al. 2011; Duc et al. 2013; Romine et al. 2014), nonetheless hinders the performance of its probabilistic forecasts. Therefore, addressing spread deficiency is a likely path toward improving the performance of high-resolution, single-physics, single-dynamical-core EPSs.

Acknowledgments. We thank Glen Romine, Ryan Sobash, and Kate Fossell of the NCAR Ensemble team [NCAR/Mesoscale and Microscale Meteorology Laboratory (MMM)] for their efforts to run the NCAR Ensemble, Eric Rogers (NOAA/NCEP/EMC) for preoperational NAM-3km forecasts, Yan Lou (NOAA/NCEP/EMC) for GFS forecasts, NOAA/NCEP/EMC for HRRR and SREF forecasts, the NRCS for providing SNOTEL data, the PRISM Climate Group at Oregon State University for providing PRISM data, and the University of Utah Center for High Performance Computing for computer-support services. Comments from John Horel, Court Strong, and three anonymous reviewers improved this paper. This article is based on research supported by the NOAA/National Weather Service CSTAR Program through Grants NA13NWS4680003 and NA17NWS4680001. Partial support is provided by NOAA Grant NA15OAR4590191. 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. NCAR is sponsored by the National Science Foundation.

REFERENCES

Alexander, C., S. S. Weygandt, S. G. Benjamin, T. G. Smirnova, J. M. Brown, P. Hofmann, and E. P. James, 2011: The High Resolution Rapid Refresh (HRRR): Recent and future enhancements, time-lagged ensembling, and 2010 forecast evaluation activities. *24th Conf. on Weather and Forecasting/20th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 12B.2, https://ams.confex.com/ams/91Annual/webprogram/Paper183065.html.

——, and Coauthors, 2014: The High-Resolution Rapid Refresh (HRRR): A maturation of frequently updating convection-allowing numerical weather prediction. *Extended Abstracts*, *World Weather Open Science Conf.*, Montreal, QC, Canada, CIRES/ESRL/NCEP, https://www.wmo.int/pages/prog/arep/wwrp/new/wwosc/documents/WWOSC2014_Alexander_Final.pdf.

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.

——, 2003: A local least squares framework for ensemble filtering. *Mon. Wea. Rev.*, **131**, 634–642, https://doi.org/10.1175/1520-0493(2003)131<0634:ALLSFF>2.0.CO;2.

Barrett, A. I., S. L. Gray, D. J. Kirshbaum, N. M. Roberts, D. M. Schultz, and J. G. Fairman, 2016: The utility of convection-permitting ensembles for the prediction of stationary convective bands. *Mon. Wea. Rev.*, **144**, 1093–1114, https://doi.org/10.1175/MWR-D-15-0148.1.

Ben Bouallègue, Z., S. E. Theis, and C. Gebhardt, 2013: Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteor. Z.*, **22**, 49–59, https://doi.org/10.1127/0941-2948/2013/0374.

Bouttier, F., B. Vié, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Mon. Wea. Rev.*, **140**, 3706–3721, https://doi.org/10.1175/MWR-D-12-00031.1.

Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 703–722, https://doi.org/10.1002/qj.234.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brill, K. F., 2009: A general analytic method for assessing sensitivity to bias of performance measures for dichotomous forecasts. *Wea. Forecasting*, **24**, 307–318, https://doi.org/10.1175/2008WAF2222144.1.

Brocker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, https://doi.org/10.1175/WAF993.1.

Bytheway, J. L., and C. D. Kummerow, 2015: Toward an object-based assessment of high-resolution forecasts of long-lived

convective precipitation in the central U.S. *J. Adv. Model. Earth Syst.*, **7**, 1248–1264, https://doi.org/10.1002/2015MS000497.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model description and implementation. *Mon. Wea. Rev.*, **129**, 569–585, https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.

Clark, A. J., 2017: Generation of ensemble mean precipitation forecasts from convection-allowing ensembles. *Wea. Forecasting*, **32**, 1569–1583, https://doi.org/10.1175/WAF-D-16-0199.1.

——, W. A. Gallus Jr., M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, https://doi.org/10.1175/2009WAF2222222.1.

——, and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, https://doi.org/10.1175/2010MWR3624.1.

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.

Colle, B. A., 2004: Sensitivity of orographic precipitation to changing ambient conditions and terrain geometries: An idealized modeling perspective. *J. Atmos. Sci.*, **61**, 588–606, https://doi.org/10.1175/1520-0469(2004)061<0588:SOOPTC>2.0.CO;2.

——, M. F. Garvert, J. B. Wolfe, C. F. Mass, and C. P. Woods, 2005: The 13–14 December 2001 IMPROVE-2 event. Part III: Simulated microphysical budgets and sensitivity studies. *J. Atmos. Sci.*, **62**, 3535–3558, https://doi.org/10.1175/JAS3552.1.

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.

——, M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031–2064, https://doi.org/10.1002/joc.1688.

Dey, S. R., 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.

Di Luzio, M., G. L. Johnson, C. Daly, J. K. Eischeid, and J. G. Arnold, 2008: Constructing retrospective gridded daily precipitation and temperature datasets for the conterminous United States. *J. Appl. Meteor. Climatol.*, **47**, 475–497, https://doi.org/10.1175/2007JAMC1356.1.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier, and B. Yang, 2015: Short Range Ensemble Forecast (SREF) system at NCEP: Recent development and future transition. *23rd Conf. on Numerical Weather Prediction/27th Conf. on Weather Analysis and Forecasting*, Chicago, IL, Amer. Meteor. Soc., 2A.5, https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273421.html.

Duc, L., K. Saito, and H. Seko, 2013: Spatial–temporal fractions verification for high-resolution ensemble forecasts. *Tellus*, **65A**, 18 171, https://doi.org/10.3402/tellusa.v65i0.18171.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

——, 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, https://doi.org/10.1002/met.25.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 456 pp.

Fassnacht, S. R., 2004: Estimating Alter-shielded gauge snowfall undercatch, snowpack sublimation, and blowing snow transport at six sites in the coterminous USA. *Hydrol. Processes*, **18**, 3481–3492, https://doi.org/10.1002/hyp.5806.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, https://doi.org/10.1175/WAF-D-15-0134.1.

Garvert, M. F., C. P. Woods, B. A. Colle, C. F. Mass, P. V. Hobbs, M. T. Stoelinga, and J. B. Wolfe, 2005: The 13–14 December 2001 IMPROVE-2 event. Part II: Comparisons of MM5 model simulations of clouds and precipitation with observations. *J. Atmos. Sci.*, **62**, 3520–3534, https://doi.org/10.1175/JAS3551.1.

Gebhardt, C., S. E. Theis, M. Paulat, and Z. Ben Bouallègue, 2011: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.*, **100**, 168–177, https://doi.org/10.1016/j.atmosres.2010.12.008.

Grubišić, V., R. K. Vellore, and A. W. Huggins, 2005: Quantitative precipitation forecasting of wintertime storms in the Sierra Nevada: Sensitivity to the microphysical parameterization and horizontal resolution. *Mon. Wea. Rev.*, **133**, 2834–2859, https://doi.org/10.1175/MWR3004.1.

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The Met Office convective-scale ensemble, MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **143**, 2846–2861, https://doi.org/10.1002/qj.3135.

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.

Hart, K. A., W. J. Steenburgh, and D. J. Onton, 2005: Model forecast improvements with decreased horizontal grid spacing over finescale intermountain orography during the 2002 Olympic Winter Games. *Wea. Forecasting*, **20**, 558–576, https://doi.org/10.1175/WAF865.1.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, https://doi.org/10.1029/2008JD009944.

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.

Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

——, 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp., http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.

Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC storm-scale ensemble of opportunity: Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *26th Conf. on Severe Local Storms*, Nashville,

TN, Amer. Meteor. Soc., 137, https://ams.confex.com/ams/26SLS/webprogram/Paper211729.html.

——, C. J. Melick, and S. J. Weiss, 2016: Comparison of the SPC storm-scale ensemble of opportunity to other convection-allowing ensembles for severe weather forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 102, https://ams.confex.com/ams/28SLS/webprogram/Paper300910.html.

Johnson, A., and X. Wang, 2016: A study of multiscale initial condition perturbation methods for convection-permitting ensemble forecasts. *Mon. Wea. Rev.*, **144**, 2579–2604, https://doi.org/10.1175/MWR-D-16-0056.1.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

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: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

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.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Melhauser, C., F. Zhang, Y. Weng, Y. Jin, H. Jin, and Q. Zhao, 2017: A multiple-model convection-permitting ensemble examination of the probabilistic prediction of tropical cyclones: Hurricanes Sandy (2012) and Edouard (2014). *Wea. Forecasting*, **32**, 665–688, https://doi.org/10.1175/WAF-D-16-0082.1.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys. Space Phys.*, **20**, 851–875, https://doi.org/10.1029/RG020i004p00851.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, https://doi.org/10.1029/97JD00237.

Munsell, E. B., J. A. Sippel, S. A. Braun, Y. Weng, and F. Zhang, 2015: Dynamics and predictability of Hurricane Nadine (2012) evaluated through convection-permitting ensemble analysis and forecasts. *Mon. Wea. Rev.*, **143**, 4514–4532, https://doi.org/10.1175/MWR-D-14-00358.1.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Neiman, P. J., F. M. Ralph, A. B. White, D. E. Kingsmill, and P. O. Persson, 2002: The statistical relationship between upslope flow and rainfall in California's coastal mountains: Observations during CALJET. *Mon. Wea. Rev.*, **130**, 1468–1492, https://doi.org/10.1175/1520-0493(2002)130<1468:TSRBUF>2.0.CO;2.

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Wea. Forecasting*, **29**, 489–504, https://doi.org/10.1175/WAF-D-13-00066.1.

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, 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.

Roe, G. H., 2005: Orographic precipitation. *Annu. Rev. Earth Planet. Sci.*, **33**, 645–671, https://doi.org/10.1146/annurev.earth.33.092203.122541.

Rogers, E., and Coauthors, 2017: Mesoscale modeling development at the National Centers for Environmental Prediction: Version 4 of the NAM forecast system and scenarios for the evolution to a high-resolution ensemble forecast system. *28th Conf. on Weather and Forecasting/24th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 3B.4, https://ams.confex.com/ams/97Annual/webprogram/Paper311212.html.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, https://doi.org/10.1175/MWR-D-14-00100.1.

Rotunno, R., and R. A. Houze, 2007: Lessons on orographic precipitation from the Mesoscale Alpine Programme. *Quart. J. Roy. Meteor. Soc.*, **133**, 811–830, https://doi.org/10.1002/qj.67.

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.

——, ——, and ——, 2015: The inland penetration of atmospheric rivers over western North America: A Lagrangian analysis. *Mon. Wea. Rev.*, **143**, 1924–1944, https://doi.org/10.1175/MWR-D-14-00288.1.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Schellander-Gorgas, T., Y. Wang, F. Meier, F. Weidle, C. Wittmann, and A. Kann, 2017: On the forecast skills of a convection-permitting ensemble. *Geosci. Model Dev.*, **10**, 35–56, https://doi.org/10.5194/gmd-10-35-2017.

Schwartz, C. S., 2014: Reproducing the September 2013 record-breaking rainfall over the Colorado Front Range with high-resolution WRF forecasts. *Wea. Forecasting*, **29**, 393–402, https://doi.org/10.1175/WAF-D-13-00136.1.

——, and Coauthors, 2009: Next-day convection-allowing WRF Model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, https://doi.org/10.1175/2009MWR2924.1.

——, G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, https://doi.org/10.1175/WAF-D-13-00145.1.

——, ——, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, https://doi.org/10.1175/WAF-D-15-0103.1.

Serreze, M. C., M. P. Clark, R. L. Armstrong, D. A. McGinnis, and R. S. Pulwarty, 1999: Characteristics of the western

United States snowpack telemetry (SNOTEL) data. *Water Resour. Res.*, **35**, 2145–2160, https://doi.org/10.1029/1999WR900090.

——, ——, 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.

Smith, R. B., and Coauthors, 2012: Orographic precipitation in the tropics: The Dominica Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 1567–1579, https://doi.org/10.1175/BAMS-D-11-00194.1.

Stoelinga, M. T., and Coauthors, 2003: Improvement of Microphysical Parameterization through Observational Verification Experiment. *Bull. Amer. Meteor. Soc.*, **84**, 1807–1826, https://doi.org/10.1175/BAMS-84-12-1807.

Tegen, I., P. Hollrig, M. Chin, I. Fung, D. Jacob, and J. Penner, 1997: Contribution of different aerosol species to the global aerosol extinction optical thickness: Estimates from model results. *J. Geophys. Res.*, **102**, 23 895–23 915, https://doi.org/10.1029/97JD01864.

Tennant, W., 2015: Improving initial condition perturbations for MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **141**, 2324–2336, https://doi.org/10.1002/qj.2524.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Wea. Rev.*, **117**, 1779–1800, https://doi.org/10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Sphenson, Eds., John Wiley and Sons, 137–163.

Trier, S. B., G. S. Romine, D. A. Ahijevych, R. J. Trapp, R. S. Schumacher, M. C. Coniglio, and D. J. Stensrud, 2015: Mesoscale thermodynamic influences on convection initiation near a surface dryline in a convection-permitting ensemble. *Mon. Wea. Rev.*, **143**, 3726–3753, https://doi.org/10.1175/MWR-D-15-0133.1.

Vié, B., G. Molinié, O. Nuissier, B. Vincendon, V. Ducrocq, F. Bouttier, and E. Richard, 2012: Hydro-meteorological evaluation of a convection-permitting ensemble prediction system for Mediterranean heavy precipitating events. *Nat. Hazards Earth Syst. Sci.*, **12**, 2631–2645, https://doi.org/10.5194/nhess-12-2631-2012.

Weisman, M. L., C. A. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, https://doi.org/10.1175/2007WAF2007005.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

Yang, D., B. E. Goodison, J. R. Metcalfe, V. S. Golubev, R. Bates, T. Pangburn, and C. L. Hanson, 1998: Accuracy of NWS 8'' standard nonrecording precipitation gauge: Results and application of WMO intercomparison. *J. Atmos. Oceanic Technol.*, **15**, 54–68, https://doi.org/10.1175/1520-0426(1998)015<0054:AONSNP>2.0.CO;2.

Zhang, F., and Y. Weng, 2015: Predicting hurricane intensity and associated hazards: A five-year real-time forecast experiment with assimilation of airborne Doppler radar observations. *Bull. Amer. Meteor. Soc.*, **96**, 25–33, https://doi.org/10.1175/BAMS-D-13-00231.1.