GEFS Precipitation Forecasts and the Implications of Statistical Downscaling over the Western United States

WYNDAM R. LEWIS AND W. JAMES STEENBURGH

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

TREVOR I. ALCOTT

NOAA/Earth System Research Laboratory, Boulder, Colorado

JONATHAN J. RUTZ

NOAA/NWS/Western Region Headquarters, Salt Lake City, Utah

(Manuscript received 11 October 2016, in final form 3 February 2017)

ABSTRACT

Contemporary operational medium-range ensemble modeling systems produce quantitative precipitation forecasts (QPFs) that provide guidance for weather forecasters, yet lack sufficient resolution to adequately resolve orographic influences on precipitation. In this study, cool-season (October–March) Global Ensemble Forecast System (GEFS) QPFs are verified using daily (24 h) Snow Telemetry (SNOTEL) observations over the western United States, which tend to be located at upper elevations where the orographic enhancement of precipitation is pronounced. Results indicate widespread dry biases, which reflect the infrequent production of larger 24-h precipitation events ($\gtrsim$22.9 mm in Pacific ranges and $\gtrsim$10.2 mm in the interior ranges) compared with observed. Performance metrics, such as equitable threat score (ETS), hit rate, and false alarm ratio, generally worsen from the coast toward the interior. Probabilistic QPFs exhibit low reliability, and the ensemble spread captures only ~30% of upper-quartile events at day 5. In an effort to improve QPFs without exacerbating computing demands, statistical downscaling is explored based on high-resolution climatological precipitation analyses from the Parameter-Elevation Regressions on Independent Slopes Model (PRISM), an approach frequently used by operational forecasters. Such downscaling improves model biases, ETSs, and hit rates. However, 47% of downscaled QPFs for upper-quartile events are false alarms at day 1, and the ensemble spread captures only 56% of the upper-quartile events at day 5. These results should help forecasters and hydrologists understand the capabilities and limitations of GEFS forecasts and statistical downscaling over the western United States and other regions of complex terrain.

1. Introduction

Accurate quantitative precipitation forecasts (QPFs) in mountainous regions are particularly challenging for meteorologists using current operational ensemble prediction systems, which lack sufficient resolution to adequately resolve critical convective and orographic processes that strongly influence the distribution and intensity of precipitation (Junker et al. 1992; Kunz and Kottmeier 2006; Smith et al. 2010; Van Haren et al. 2015). Over the western United States, for example, meteorologists must infer how local terrain features will modulate rainfall and snowfall, as well as precipitation impacts on air and ground transportation, water resource and flood management, outdoor recreation, and avalanche safety (Stewart et al. 1995; Cohen 1996; Ralph et al. 2006; Neiman et al. 2011; U.S. Department of the Interior 2012; Black and Mote 2015; Schirmer and Jamieson 2015; Parker and Abatzoglou 2016). Knowledge of model biases, capabilities, and limitations has the potential to improve forecasts, but is limited by a scarcity of studies evaluating operational ensemble model performance in areas of complex terrain (e.g., Schirmer and Jamieson 2015).

The majority of western U.S. precipitation occurs during the cool season, defined here as October–March, with a significant portion falling as snow at higher elevations (Serreze et al. 1999). Large precipitation events with

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

DOI: 10.1175/WAF-D-16-0179.1

high snow levels, many associated with atmospheric rivers (ARs), yield hydrological extremes that produce flooding, property and infrastructure damage, and loss of life (Ralph et al. 2006; Neiman et al. 2011; U.S. Department of the Interior 2012; Rutz et al. 2014). For example, orographic enhancement during AR conditions produced all seven major floods on California's Russian River from October 1997 to February 2006 (Ralph et al. 2006). Many mountain areas and highways in the western United States are also susceptible to avalanche hazards, with snowfall and rainfall increasing the likelihood of natural and human-triggered avalanches (Tremper 2008; Hatchett et al. 2017). State Route 210 in Utah, for example, crosses 50 avalanche paths and is hit by an average of 33 avalanches per year (Steenburgh 2014). Winter precipitation–related motor vehicle and aviation accidents result in roughly 900 fatalities on average each year across the United States, with some of the highest standardized mortality rates occurring in the West (Black and Mote 2015). Nationally, such fatalities amount to more than double the combined fatalities from lightning, tornadoes, hurricanes, heat, and cold (Black and Mote 2015).

QPFs are typically more skillful in the cool season when large-scale dynamic forcing dominates precipitation generation, as opposed to the localized convection and weaker dynamic forcing found during the warm season (Junker et al. 1992; Mullen and Buizza 2001; Baxter et al. 2014). Nevertheless, QPF skill is often lower in mountainous regions, due at least in part to poorly resolved terrain features (Junker et al. 1992; Yuan et al. 2005; Ikeda et al. 2010) and, over the western U.S. interior, low spatial coherence of precipitation events (Serreze et al. 2001; Parker and Abatzoglou 2016). Complex terrain also contributes to cyclone displacement errors in numerical forecasts (Charles and Colle 2009), which in turn affects the position and timing of precipitation features, as well as moisture transport associated with atmospheric rivers (e.g., Rutz et al. 2014, 2015).

This study focuses on the Global Ensemble Forecast System (GEFS), an operational ensemble modeling system run by the National Weather Service (NWS) that is widely used by forecasters in the western United States. With an effective horizontal grid spacing of ~33 km, the GEFS is unable to resolve key topographical features and subsequent effects on precipitation (NOAA 2015). While we are unaware of any peer-reviewed analyses examining the performance of the current version of the GEFS, which became operational in December 2015, Hamill (2012) showed that an earlier version of the GEFS produced probabilistic QPFs (PQPFs) with insufficient spread, lower reliability, and lower Brier skill scores (BSSs) compared with the European, Canadian, and U.K. ensemble modeling systems. Baxter et al. (2014) also evaluated an earlier version of the GEFS, showing that GEFS QPFs have little useful skill over the southeast United States by forecast day 5.5 (108–132 h), and that GEFS PQPFs demonstrate little to no skill compared with climatological event frequencies by forecast day 6.5 (132–156 h).

Approaches aimed at improving QPFs or PQPFs from coarse-resolution ensembles include ensemble-MOS approaches (see Wilks 2006a for a review), calibration using rank histograms (Hamill and Colucci 1997, 1998; Eckel and Walters 1998), dynamic downscaling (e.g., Stensrud et al. 1999; Marsigli et al. 2001), bias correction and statistical disaggregation (Wood et al. 2002), Bayesian model averaging (Raftery et al. 2005; Sloughter et al. 2007; Fraley et al. 2010; Schmeits and Kok 2010), analog sorting (Bontron and Obled 2005), reforecast analogs (Hamill and Whitaker 2006; Hamill et al. 2015), and fitting to censored, shifted gamma distributions (Scheuerer and Hamill 2015). In this study, we seek to identify the capabilities and limitations of statistical downscaling based on climatological precipitation analyses because it is computationally inexpensive, widely employed in climate and hydrological applications (e.g., Wilby et al. 1998; Wood et al. 2004; Gutmann et al. 2012), and used by the Weather Prediction Center (WPC) and many NWS Forecast Offices and River Forecast Centers in the western United States.

For these operational applications, such downscaling typically uses high-resolution (~4-km or ~800-m grid spacing) precipitation analyses produced by the PRISM Climate Group at Oregon State University (Daly et al. 1994, 2008) to rescale lower-resolution model guidance and provide increased spatial detail. Described in greater depth in section 2b, the approach implicitly assumes climatological precipitation distributions and that the small-scale precipitation variability is directly related to the large-scale precipitation pattern. This yields climatologically plausible precipitation distributions, but may be problematic during storms that are strongly influenced by unresolved mesoscale processes (e.g., mesoscale precipitation bands, nonorographic convection, etc.) or feature precipitation–altitude relationships that deviate from climatology (e.g., Steenburgh 2003, 2004), particularly in regions where orographic enhancement is sensitive to flow direction.

The purpose of this study is to provide a comprehensive overview of GEFS QPF and PQPF performance in upper-elevation regions of the western United States, including an evaluation of statistical downscaling using high-resolution PRISM climatology. Specifically, we examine the performance of the current operational version of the GEFS relative to the NOAA/Climate Prediction Center (CPC) Unified Daily Precipitation Analysis (hereafter the CPC analysis) and upper-elevation Snow Telemetry (SNOTEL) observations. These datasets and the methods used for evaluation are described in section 2. Results are then

presented in section 3, with conclusions and a discussion of the significance of our findings provided in section 4.

2. Data and methods

a. Global Ensemble Forecast System

We verify reforecasts (i.e., retrospective forecasts) and forecasts produced by the current (as of 2 December 2015) operational version of the GEFS,[1] which is based on version 12.1.0 of the National Centers for Environmental Prediction (NCEP) Global Spectral Model [the forecast component of the Global Forecast System (GFS)], configured with 64 vertical levels and a horizontal resolution of TL574 (~33 km) for the first 192 h and TL382 (~55 km) from 192 to 384 h (NOAA 2015). Ensemble members consist of a control and 20 perturbations generated with an ensemble Kalman filter scheme (Wang et al. 2013; Hou et al. 2015). Reforecasts for the 2013–14 and 2014–15 cool seasons were obtained from the NOAA Operational Model Archive and Distribution System (NOMADS) data server (Rutledge et al. 2006), whereas reforecasts (1 October–1 December 2015) and forecasts for the remainder of the 2015–16 cool season were provided by NCEP's Environmental Modeling Center. Although GEFS forecasts are currently available four times a day on a 0.5° latitude–longitude grid, we use 0000 UTC initialized runs on a 1.0° latitude–longitude grid since this is the only initialization time and output grid spacing available on the NCEP NOMADS server for the 2013–14 and 2014–15 reforecast periods. Given that the GEFS is typically available a few hours after the nominal initialization time, we define day 1 as the 12–36-h forecast and perform validation through day 7 (156–180 h), which concentrates on the higher-resolution portion of the GEFS forecasts.

b. Downscaling methodology

The climatology-based statistical downscaling method used here is similar to an algorithm used frequently by NWS meteorologists to downscale coarse-resolution model QPFs during the preparation of graphical forecasts for the National Digital Forecast Database. Such downscaling uses monthly, climatological (1981–2010) high-resolution (30 arc s, ~800-m grid spacing) precipitation analyses produced by the PRISM Climate Group at Oregon State University [analysis technique described by Daly et al. (1994)]. First, we generate a daily precipitation climatology for the forecast day of interest by assuming monthly PRISM values are valid on the 15th of each month and then interpolating to daily values (Fig. 1a). We then smooth the daily values to a spatial scale approximately consistent with the GEFS 1.0° latitude–longitude grid (Fig. 1b). In operational practice, a variety of techniques are used for this smoothing including Gaussian filtering (K. Brill, WPC, 2016, personal communication) and area averaging (T. Barker, NWSFO Boise, 2016, personal communication). We use an approach similar to that of the WPC with a Gaussian filter with $\sigma = 0.5°$, which yielded the smallest biases compared with area-average approaches or Gaussian filters with different σ values. Next, we divide the original PRISM precipitation analysis by the Gaussian-filtered analysis to obtain an analysis of the *downscaling ratio* across the western United States (Fig. 1c), imposing a lower bound of 0.3 and an upper bound of 5 to avoid extreme outliers, although these thresholds are rarely met.

Bilinearly interpolating the GEFS QPF (Fig. 1d) to the PRISM grid (Fig. 1e) and multiplying by the downscaling ratio yields the downscaled QPF (Fig. 1f). The downscaling ratio is typically less than 1 in valleys and basins, leading to a downscaled QPF that is lower than the GEFS QPF. Conversely, the downscaling ratio is typically greater than 1 in mountains and upland regions, leading to a downscaled QPF that is larger than the GEFS QPF. For point verification in this study, GEFS QPFs and daily downscaling ratios are bilinearly interpolated to observation locations and multiplied to obtain downscaled QPFs.

c. Precipitation analyses and observations

To identify regional biases in the GEFS reforecasts and forecasts, we use the CPC analysis on a 0.25° latitude–longitude grid (Higgins et al. 2000; Xie et al. 2007; Chen et al. 2008) and bilinearly interpolate GEFS QPFs to the CPC analysis grid for comparison. Although higher-resolution precipitation analyses are available [e.g., the Climatology-Calibrated Precipitation Analysis (Hou et al. 2014)], the lower-resolution CPC analysis is sufficient for identifying broad regional biases in GEFS forecasts.

Gauge-based verification in upper-elevation regions uses accumulated [since 0000 Pacific standard time (PST) 1 October] precipitation observations from the SNOTEL network maintained by the National Resources Conservation Service (NRCS). The automated SNOTEL stations measure precipitation collected by a large-storage weighing gauge in imperial units at 0.1-in. (~2.5 mm) precision. SNOTEL stations are typically placed in sheltered areas with regionally high snow accumulations and include an Alter wind shield to reduce undercatch (Yang et al. 1998; Serreze et al. 1999; Fassnacht 2004). Comparable gauges have shown an undercatch of ~10%–15%

[1] The reforecasts were generated by NCEP and should not be confused with those from the NOAA/Earth System Research Laboratory Physical Sciences Division second-generation reforecast project (Hamill et al. 2013), which uses an older version of the GEFS run at lower resolution.

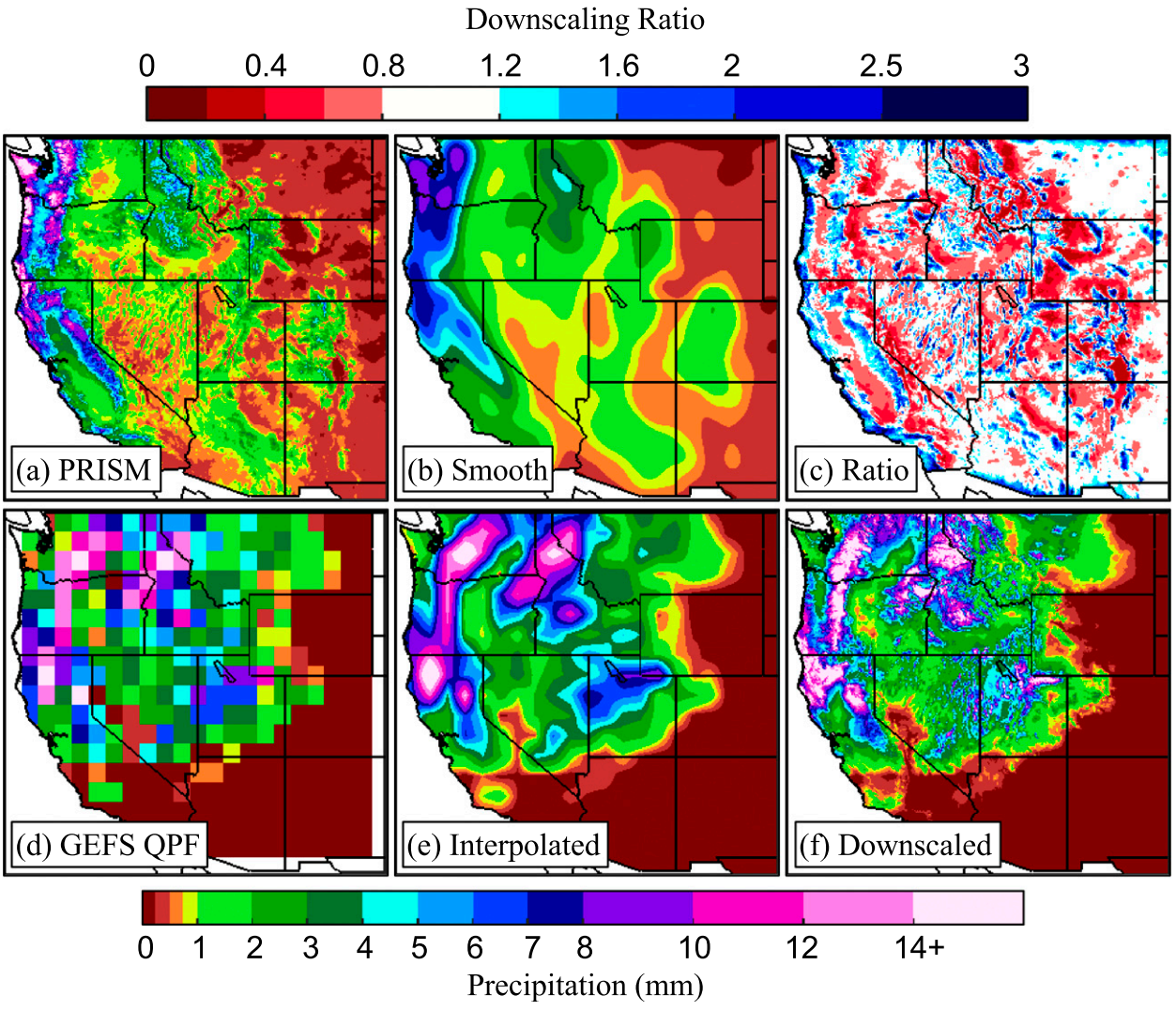

FIG. 1. Statistical downscaling example. (a) The 24 Jan 2016 PRISM climatological precipitation. (b) As in (a), but smoothed with a Gaussian filter. (c) Downscaling ratio derived by dividing (a) by (b). (d) The 24 Jan 2016 GEFS day 1 control forecast as provided by NCEP. (e) As in (d), but bilinearly interpolated onto a PRISM lat–lon grid. (f) As in (d), but downscaled by multiplying (c) and (e).

for wind speeds of about 1–2 m s^{-1} (Yang et al. 1998; Fassnacht 2004; Rasmussen et al. 2012), which is a typical wind speed found in forest clearings that house SNOTEL stations (Ikeda et al. 2010). Additional factors influencing SNOTEL precipitation data include transmission errors, instrument malfunction (e.g., leaks), temperature-based fluctuations (affecting readings by the pressure transducer), and snow adhesion to gauge walls (delaying precipitation measurement). See Serreze et al. (1999) and Avanzi et al. (2014) for summaries of the capabilities and limitations of SNOTEL measurements.

Instrument limitations warrant our implementation of basic quality control to reduce the use of erroneous data. We begin with hourly cumulative precipitation observations downloaded from the NRCS, identifying negative values (typically −99.9 and −0.1 in.). If these values are surrounded by equal nonnegative values, we replace the negative values with the surrounding nonnegative value; otherwise, they are flagged as erroneous. We also adjust positive values surrounded by equal positive values to the surrounding positive value, which results in a smoother hourly time series. We then discretely sample the 1200 UTC observations and flag spikes of more than (less than) 0.5 in. above (below) the maximum (minimum) of the surrounding 20 days. Any flagged data surrounded by equal values are replaced with the equal value. Then, as done with the hourly data, we adjust positive values surrounded by equal positive values to the surrounding positive value, which results in a smoother daily time series and was a key step in the quality control approach of

Serreze et al. (1999). After these adjustments, we calculate the daily (1200–1200 UTC) precipitation for all periods when data are available for the current and prior day, setting all negative values to zero and flagging all values in excess of 5.0 in. as erroneous. The latter removes many false jumps in the data, along with a small sample of actual extreme events. The number of these extreme events is, however, small. Daily precipitation values valid at 1200 UTC 1 October require accumulated precipitation data from the previous water year and are not included.

After these checks, we remove stations that contain erroneous data on 20% of the days during the three cool seasons. Then, for each remaining station, we calculate the ratio of cumulative daily precipitation during the three cool seasons relative to that obtained from the gauge's accumulated measurement at the end of each cool season. We then remove stations at which this ratio is 1 (1.5) standard deviation above (below) the median ratio for all stations. The more relaxed criterion for lower ratios reflects the removal of events > 5 in. from the daily precipitation. Daily precipitation data are then converted from inches to millimeters. These requirements result in data from 603 of 781 stations being used for the validation.

d. Verification methods

No single statistical measure can adequately diagnose strengths and weaknesses of a numerical forecast system (Schaefer 1990). We use a series of measures based on a 2×2 contingency table commonly used for precipitation validation (Table 1), to provide a broad assessment of the capabilities of the GEFS and downscaled GEFS. These measures are described in Mason (2003) and include

$$\text{Hit rate} = \frac{a}{a+c},$$

$$\text{False alarm ratio} = \frac{b}{a+b},$$

$$\text{Bias score} = \frac{a+b}{a+c},$$

and

$$\text{ETS} = \frac{a-a_r}{a-a_r+c+b},$$

where

$$a_r = \frac{(a+c)(a+b)}{n}.$$

The hit rate is equal to the fraction of correct forecasts (hits) to observed events. The false alarm ratio expresses the fraction of forecasts that do not verify as events. The bias score represents the fraction of forecasts issued to events observed. The equitable threat score (ETS) is a common precipitation verification tool for two-category (dichotomous) events, providing a single value between 1 (perfect forecast) and 0 (equivalent to a random forecast) (Mason 2003; Hamill and Juras 2006). Varying climatological event frequencies among stations can affect ETS and other performance metrics. Following Hamill and Juras (2006), we attempted to account for this by calculating ETS as a weighted average of ETS for 10 subgroups of SNOTEL stations with similar climatological event frequencies, but we found this approach yielded results similar to traditional ETS calculations. Therefore, we use the traditional ETS.

TABLE 1. Contingency table used for forecast validation.

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

The probabilistic verification utilizes reliability diagrams [illustrating the relation of forecast probabilities to observed frequencies; Hamill (1997)], BSS [a measure of probabilistic forecast skill relative to climatological event frequencies; Brier (1950)], rank histograms [indicating where observations fall within the ensemble spread; Hamill (2001)], and additional forecast attributes to help gauge the overall value of the GEFS (Toth et al. 2003). To account for variations in climatological event frequencies across stations (Wilks 2006b, chapter 7; Hamill et al. 2008), reliability diagrams include a histogram inset that displays the frequency of occurrence of forecast probabilities and the SNOTEL climatological event frequencies in 10% bins. We also use resampling to generate 5% and 95% consistency bars, which indicate the variability among observed frequencies due to limited counting statistics (Toth et al. 2003; Brocker and Smith 2007). The approach is similar to the bootstrapping methods in Hamill et al. (2008) and follows a technique known as consistency resampling (Brocker and Smith 2007). We resample 1000 times, using N^2 samples, where N is the number of samples in each forecast probability bin. See Brocker and Smith (2007) for details.

3. Results

a. GEFS climatology

We begin by comparing mean-daily precipitation in the CPC analysis with that produced by the GEFS day 1 control forecast to describe the climate of the three cool-season study periods and identify regional-scale climatological biases in the GEFS (biases and other forecast

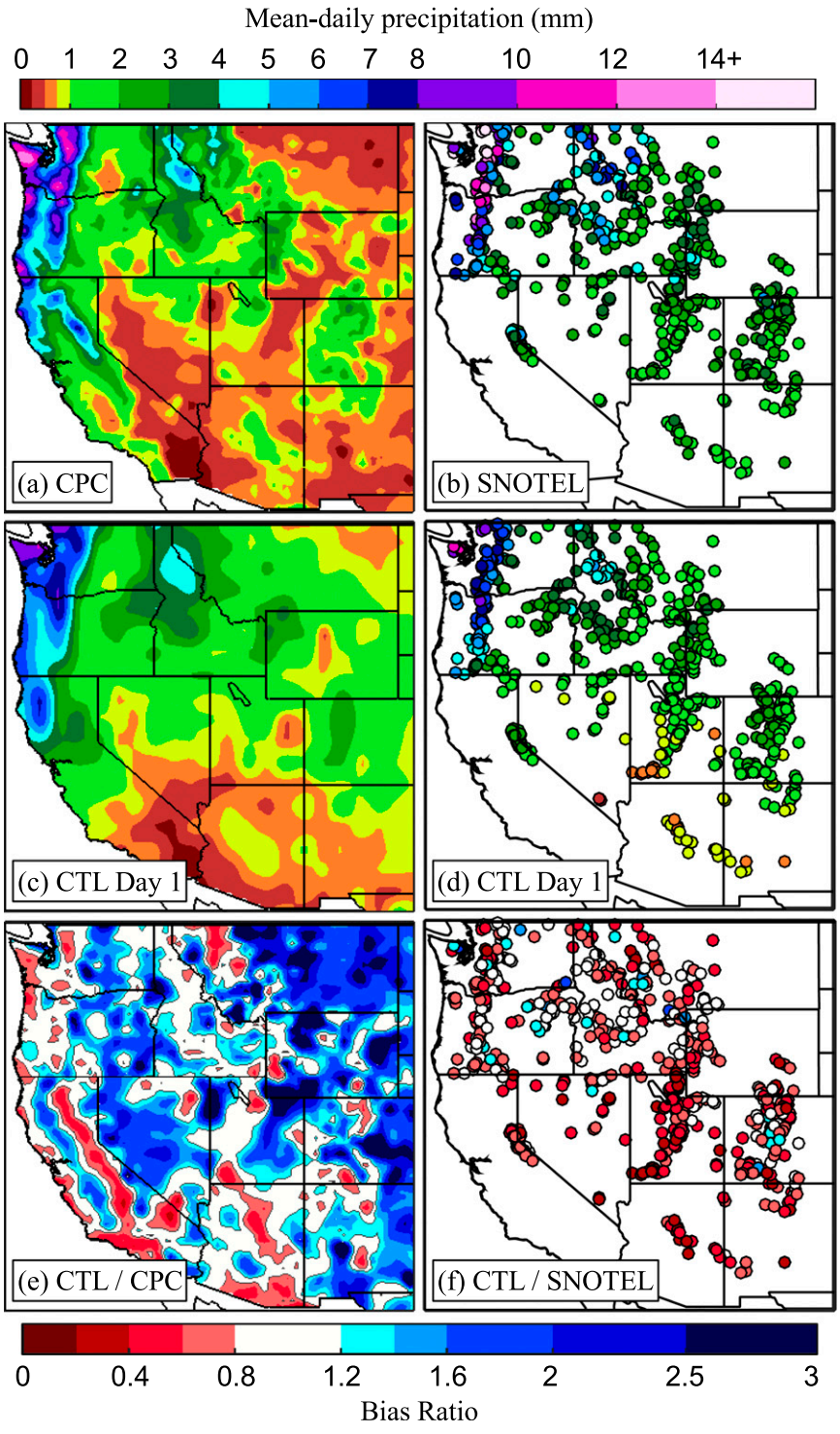

FIG. 2. Mean daily precipitation (mm; top scale) from the (a) CPC analysis, (b) SNOTEL observations, (c) GEFS day 1 (12–36 h) control forecast (CTL) interpolated to the CPC grid, and (d) GEFS day 1 CTL interpolated to SNOTEL stations. (e) GEFS day 1 CTL bias ratio (bottom scale) relative to the CPC analysis. (f) As in (e), but relative to SNOTEL observations.

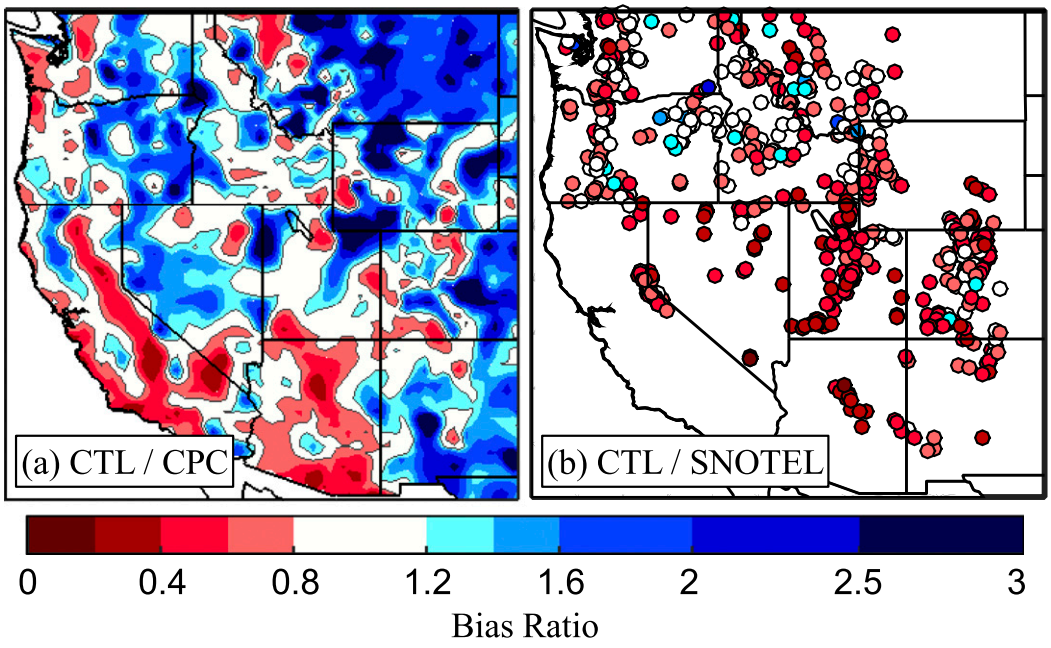

FIG. 3. As in Figs. 2e,f, but for day 5 (108–132 h).

characteristics exhibited by the GEFS control are similar to those of the other individual GEFS members). During the three-cool-season study period, CPC-analyzed precipitation was heaviest in the coastal ranges of the Pacific Northwest and northern California and the Cascade Mountains of Washington and Oregon (Fig. 2a). Over the interior, precipitation was heaviest over regions with higher terrain including northern and central Idaho, northwest Montana, north-central Utah, western Colorado, and the Mogollon Rim. The interior northwest was wetter than the interior southwest, which reflects both climatology and persistent drought conditions over the latter. The GEFS day 1 control captures the broad regional characteristics of the CPC precipitation distribution (Fig. 2c); however, the ratio of GEFS control to CPC precipitation (i.e., the bias ratio) reveals that the GEFS control is too dry over and upstream of topographic barriers and too wet in downstream valleys and basins (Fig. 2e). When comparing the GEFS control at SNOTEL stations (Fig. 2d) to SNOTEL observations (Fig. 2b), a dry bias is evident at most stations (Fig. 2f), which are located preferentially in upper-elevation regions. At 22% (60%) of the SNOTEL stations, the bias ratio is smaller than 0.5 (0.75), indicating a substantial dry bias. By day 5, the GEFS control bias ratio relative to both the CPC analysis and SNOTEL stations has shifted to slightly lower values, revealing a tendency for the GEFS control (as well as other individual GEFS members) to become drier with increasing forecast lead time (cf. Figs. 2e, 3a and 2f, 3b).[2] Specifically, the GEFS control produced 5% less precipitation at day 5 compared with day 1 at SNOTEL stations and across the western United States as a whole.

A comparison of the frequency of daily (24 h) precipitation (2.54-mm bins) produced by the GEFS control with that at CPC analysis grid points (Fig. 4a) and SNOTEL stations (Fig. 4b) identifies biases in event frequency as a function of event size. Compared to the CPC analysis, which spans the low and high elevations of the western United States, the GEFS day 1 control produces too many events $\lesssim$ 20.3 mm and too few events $\gtrsim$ 22.9 mm (Fig. 4a, frequency bias indicated by the black dotted line and right ordinate). The largest frequency bias (forecast frequency/observed frequency) is associated with 5.1-mm events, above which the frequency bias exhibits a near-monotonic decline with increasing event size (Fig. 4a). For all western U.S. SNOTEL stations, events $\lesssim$ 7.6 mm are predicted at a frequency consistent with the observations, while events $\gtrsim$ 10.2 mm are associated with an underprediction of event frequency that worsens with increasing event size (Fig. 4b). Consideration of undercatch, as might be expected with SNOTEL

[2] Because there are a few days with missing GEFS forecasts, there is a small difference in the observed mean daily precipitation on days day 1 forecasts are valid compared with days day 5 forecasts are valid. For brevity, we do not present mean daily precipitation for the latter since it closely matches Figs. 2a,b.

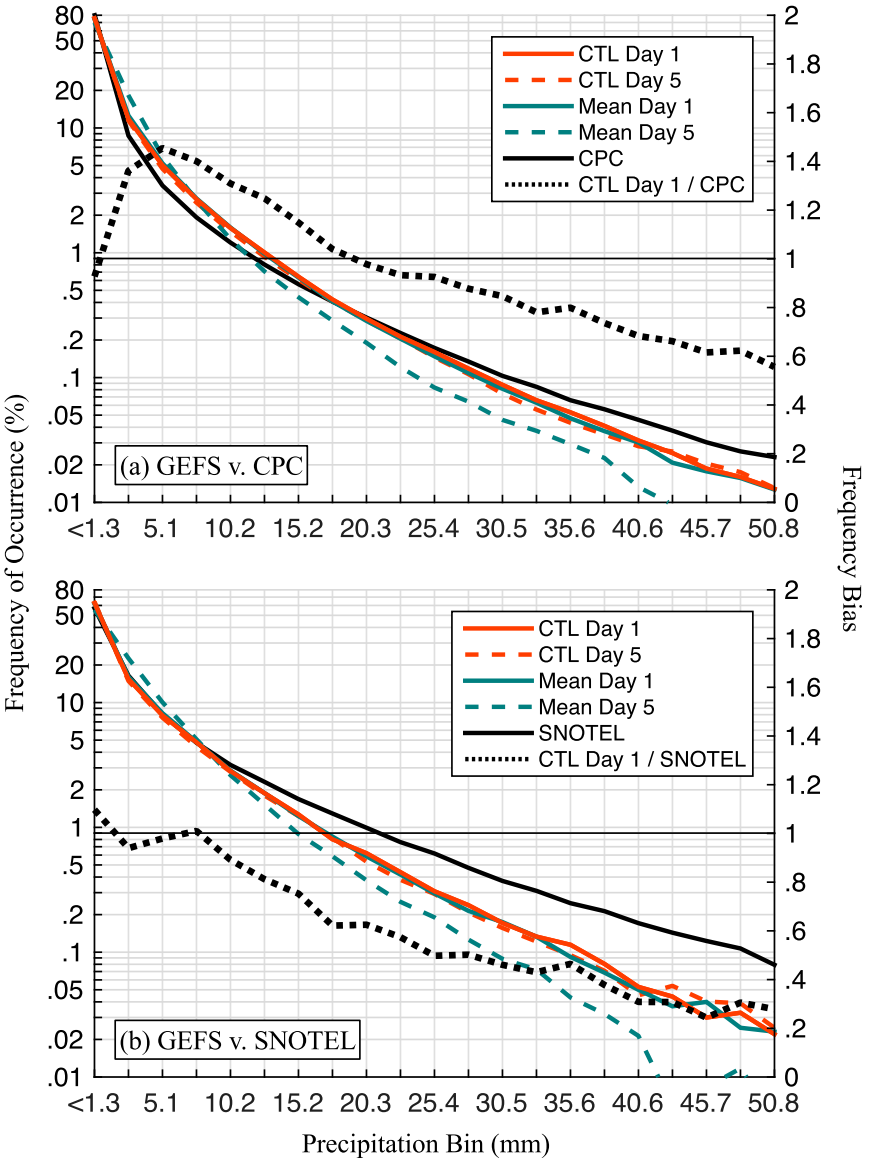

FIG. 4. (a) Precipitation frequency of the GEFS day 1 (12–36 h) CTL (CTL day 1, red line), GEFS day 5 (108–132 h) CTL (CTL day 5, dashed red line), GEFS day 1 ensemble mean forecast (mean day 1, teal line), GEFS day 5 mean (mean day 5, dashed teal line), and the CPC analysis (CPC, black line) for all CPC analysis grid points in the western United States. Bias ratio of the GEFS day 1 CTL to CPC analysis (CTL day 1/CPC) indicated by the dotted black line. (b) As in (a), but for the precipitation frequency at SNOTEL stations (SNOTEL, black line) and bias ratio of the GEFS day 1 CTL to SNOTEL observations (CTL day 1/SNOTEL, dotted black line).

gauges (Serreze et al. 1999), would further amplify the underprediction. Similar results are found for day 5 (bias ratio not shown for clarity). Averaging across all members has little impact at day 1 when the ensemble spread is small, but averaging at longer lead times results in an increased number of smaller events and a decreased number of larger events, exacerbating these frequency biases (bias ratios for ensemble means also not shown for clarity).

Distinct regional differences in frequency bias are revealed when grouping SNOTEL stations based on geography, climate, and model performance. We examined several regional groupings but ultimately present results from two highly differentiated regions: Pacific ranges and interior ranges (Fig. 5; stations from intermediate stations not presented for brevity). In the Pacific ranges, consisting of stations in the Cascade Mountains,

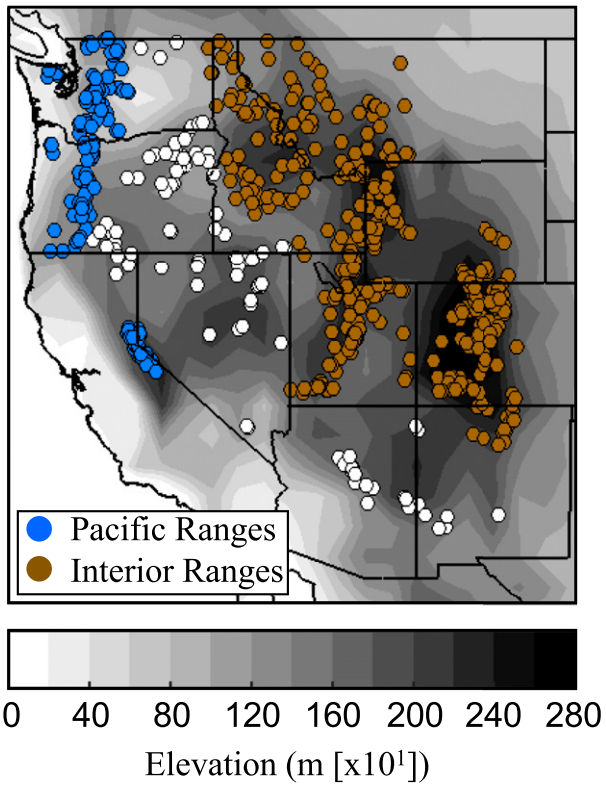

FIG. 5. Regional classification of SNOTEL stations with 1° × 1° GEFS topography (shaded following scale at bottom).

Sierra Nevada, and coastal ranges of the Pacific Northwest, the GEFS control produces similar-to-observed event frequencies (i.e., 0.8 ≤ frequency bias ≤ 1.2) for event sizes ≤ 22.9 mm (Fig. 6a). For the interior ranges, consisting of inland stations of the Pacific Northwest, Utah, and the Rocky Mountains of Wyoming, Colorado, and New Mexico, similar-to-observed event frequencies are confined to event sizes ≤ 10.2 mm (Fig. 6b). Above these thresholds, frequency biases in both regions asymptote toward zero with increasing event size, but are consistently lower in the interior ranges, reflective of a larger underprediction bias (cf. Figs. 6a,b). We hypothesize that the greater underprediction of event frequency in the interior ranges partly reflects the finescale nature of the topography and inherently low spatial coherence of precipitation systems over the western interior (Serreze et al. 2001; Parker and Abatzoglou 2016).

Bivariate histograms comparing observed and forecast precipitation provide an additional perspective on the GEFS control performance (Fig. 7). Skewness in the distribution of more frequent forecast–observation pairs relative to the 1-to-1 line confirms that at all but the smallest thresholds, observed events are more likely than not to be underforecast at day 1 in the Pacific ranges (Fig. 7a), with the underforecasting worsening over the interior ranges (Fig. 7b). To be precise, observed events ≥ 15.2 mm (5.1 mm) in the Pacific (interior) ranges at day 1 are at least twice as likely to be underforecast as overforecast, whereas events ≥ 25.4 mm (12.7 mm) are at least 5 times as likely to be underforecast. Compared with day 1, day 5 forecasts exhibit greater scatter with frequency isolines oriented more normal to the 1-to-1 line, especially in the interior ranges, which is consistent with declining skill with increasing forecast lead time (Figs. 7c,d).

b. GEFS downscaled climatology

Next, we evaluate the mean-daily precipitation from the downscaled GEFS control relative to SNOTEL observations (Fig. 8; see Fig. 2b for SNOTEL mean-daily precipitation). At days 1 and 5, downscaling addresses the widespread underprediction evident over mountains in the GEFS control, yielding wetter precipitation climatologies at 91% of SNOTEL stations. Considering all western U.S. stations, downscaling increases the median bias ratio at day 1 from 0.67 to 1.01 and at day 5 from 0.62 to 0.94. At day 1, 17%, 56% and 27% of the stations have dry (<0.8), near-neutral (0.8–1.2), and wet (>1.2) bias ratios, respectively, with a greater fraction of stations over the interior ranges exhibiting wet bias ratios (Fig. 8c). Consistent with the GEFS control becoming drier with increasing forecast lead time, the downscaled GEFS control bias ratios generally shift slightly to lower values by day 5 (Fig. 8d).

The downscaled GEFS day 1 control also demonstrates improvements over the undownscaled GEFS control for event frequency biases at SNOTEL stations. In the Pacific ranges (Fig. 9a) and the interior ranges (Fig. 9b), event frequencies are predicted reasonably well at all thresholds up to 50.8 mm, with frequency biases ranging from approximately 0.8 to 1.2.

Bivariate histograms further illustrate that events at day 1 are less likely to be underforecast by the downscaled GEFS day 1 control than the undownscaled GEFS control, with the distribution centered closer to the 1-to-1 line over both the Pacific and interior ranges, especially for larger events sizes (cf. Figs. 7a, 10a and 7c, 10c). The large departures of median observed (forecast) values below (above) the 1-to-1 line indicate that at larger event thresholds, an observed event is more likely than not to be underforecast, but, when an event is predicted, it is more likely than not to be overforecast, especially over the interior ranges. Like the undownscaled GEFS, downscaled forecasts exhibit little skill by day 5 (Figs. 10b,d).

c. Deterministic verification

Further verification of model performance focuses on upper-quartile precipitation events at CPC grid points and SNOTEL stations. Here, the upper quartile is defined as the 75th percentile of observed precipitation

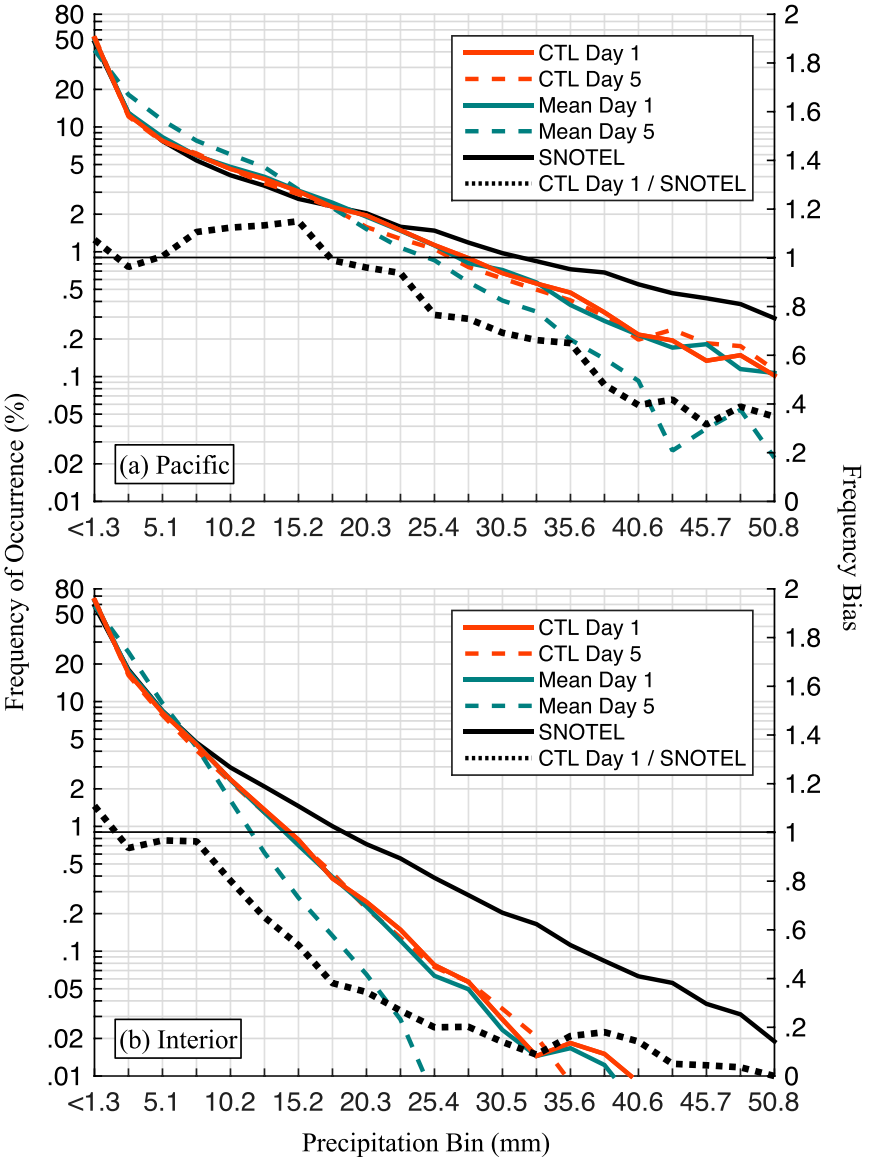

FIG. 6. As in Fig. 4b, but for (a) Pacific and (b) interior range SNOTEL stations.

events ≥ 2.54 mm (the lowest observable amount at SNOTEL stations) at each grid point or station. Aside from performance measures inevitably degrading as thresholds are increased, the spatial characteristics of the results are generally consistent for other percentile thresholds (e.g., top decile) or absolute precipitation amounts (e.g., 10 mm).

When evaluated using the CPC analysis, GEFS day 1 control ETSs are generally highest along the Pacific coast and decrease toward the interior with considerable spatial variability (Fig. 11a; other GEFS members exhibit similar performance characteristics). Compared with SNOTEL observations, GEFS day 1 control ETSs also exhibit a tendency to decline from the coastal Pacific toward the interior with considerable spatial variability (Fig. 11b). ETSs are also generally lower at sites in the interior southwest compared with the interior northwest. Downscaling of the GEFS day 1 control yields ETS improvements at 81% of SNOTEL stations in the western United States, increasing the median ETS from 0.22 to 0.34. The greatest improvements are realized over the interior, especially over Utah and Arizona (cf. Figs. 11b,c). Although ETSs do increase with downscaling over Montana and Colorado, scores remain relatively low.

Spatial patterns in ETS change minimally with increasing forecast lead time, so we instead examine cumulative statistics for upper-quartile events at all SNOTEL stations. Not surprisingly, GEFS control ETSs decline with increasing forecast lead time, dropping from 0.24 at day 1

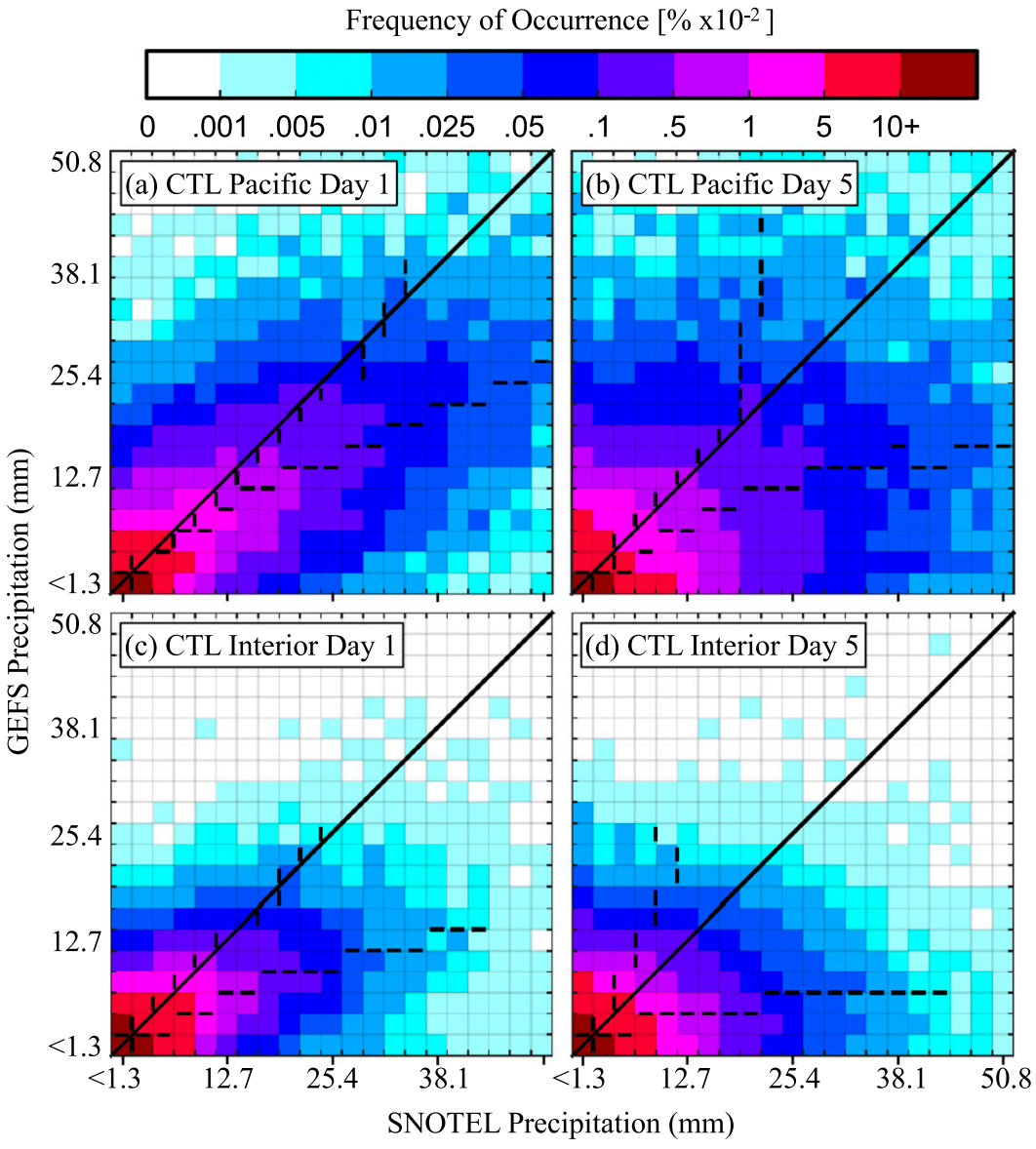

FIG. 7. Bivariate histograms of (a) GEFS day 1 (12–36 h) CTL and observed precipitation at Pacific ranges SNOTEL stations. (b) As in (a), but for GEFS day 5 (108–132 h) CTL. (c),(d) As in (a),(b), but for interior range SNOTEL stations. Horizontal (vertical) bars represent the median observed (forecast) value in each bin. Bars are not shown for bins with fewer than 100 events.

to 0.11 by day 6 (Fig. 12a), near the 0.1 threshold of useful skill identified by Baxter et al. (2014). Downscaling increases ETSs for all forecast days (Fig. 12a), with useful skill extended to day 7. Based on ETSs, the skill of the downscaled GEFS control at day 4 is approximately equivalent to the undownscaled GEFS control at day 1. ETSs for the GEFS mean (i.e., average of the control plus 20 ensemble members) are slightly worse than the control, whereas the difference between the downscaled control and downscaled GEFS mean is negligible. We suspect that the dry bias of the undownscaled GEFS results in lower ETSs for the GEFS mean compared with the control, especially at longer lead times when the ensemble spread is large.

The underprediction of larger events by the GEFS control is evident in the bias score, with values < 0.6 at all lead times (Fig. 12b). Thus, for all SNOTEL stations, the GEFS control produces about half as many upper-quartile events as observed. Downscaling substantially increases the occurrence of larger QPFs, yielding a bias score of ~1 through day 7 (Fig. 12b). Bias scores for the GEFS mean and downscaled GEFS mean are slightly

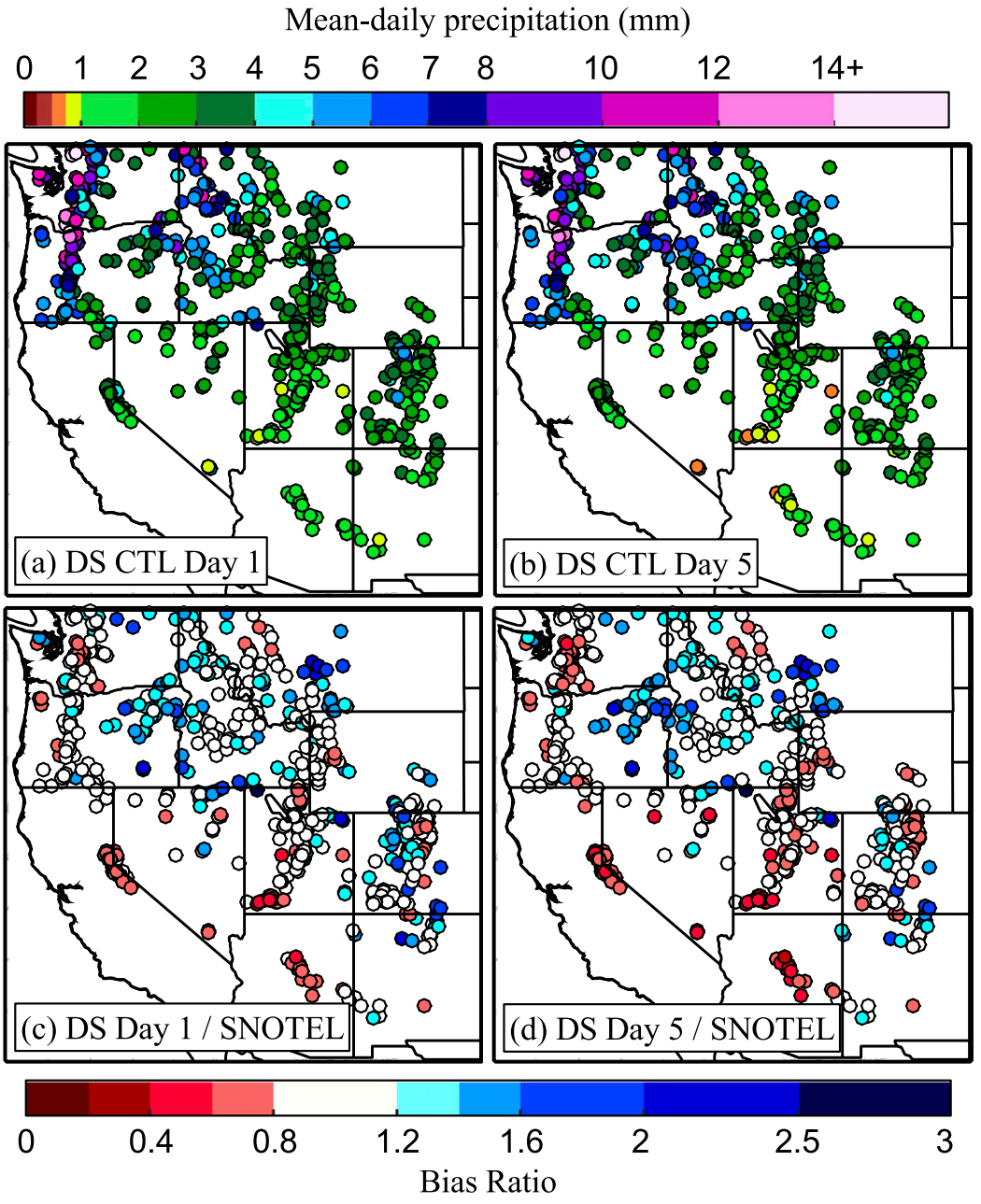

FIG. 8. Mean daily precipitation (mm; top scale) from the (a) GEFS day 1 (12–36 h) downscaled control forecast (DS CTL) at SNOTEL stations. (b) As in (a), but for day 5 (108–132 h). (c) GEFS day 1 DS CTL bias ratio (bottom scale) relative to SNOTEL observations. (d) As in (c), but for day 5.

lower than the GEFS control and downscaled GEFS control at day 1, respectively, but decline more rapidly with increasing lead time as the ensemble spread grows and averaging reduces the number of upper-quartile events forecasted (Fig. 12b).

The downscaled GEFS control exhibits a much higher hit rate than the control, although hit rates do decrease with increasing lead time, as expected (Fig. 12c). At day 1, the GEFS control upper-quartile hit rate is 0.32, with downscaling increasing this value to 0.57. However, the downscaled GEFS control also produces more false alarms than the GEFS control, with a false alarm ratio at day 1 of 0.47 that increases with forecast lead time (Fig. 12d). The GEFS mean produces hit rates and false alarm ratios analogous to the control at short lead times (Figs. 12c,d). At longer lead times, the mean produces

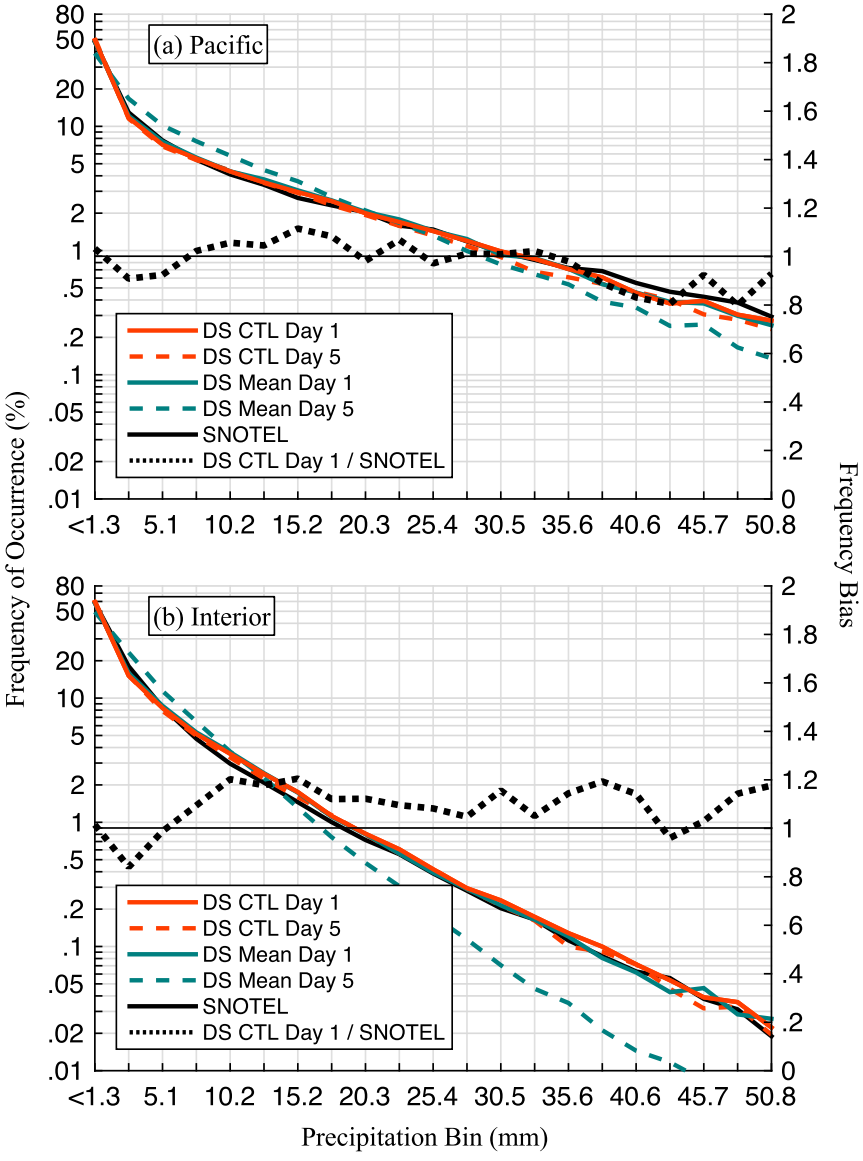

FIG. 9. As in Fig. 6, but for the downscaled GEFS.

fewer false alarms but also fewer hits than the control, as averaging an increasing ensemble spread produces fewer upper-quartile events.

Broken down by region, all four of these metrics show a decline in performance from the Pacific ranges to the interior ranges. In the Pacific ranges, ETSs and hit rates are higher (Figs. 13a,c), bias scores are closer to 1 (Fig. 13b), and false alarm ratios are lower (Fig. 13d) at all lead times. Based on ETS, a day 5 GEFS control forecast over the Pacific ranges is as skillful as a day 1 forecast over the interior ranges (Fig. 13a). ETSs for the GEFS control in the Pacific ranges are even higher than those for the downscaled GEFS control over the interior ranges at all forecast lead times, illustrating that even with downscaling, forecast performance is worse over the interior ranges than in the undownscaled GEFS control in the Pacific ranges. The GEFS day 1 control hit rate for upper-quartile events is 0.44 (0.27) over the Pacific (interior) ranges, with a false alarm ratio of 0.27 (0.41) (Figs. 13c,d). Downscaling improves day 1 hit rates to 0.61 (0.54) but worsens false alarm ratios to 0.33 (0.51).

d. Probabilistic verification

Probabilistic verification similarly concentrates on upper-quartile events. We begin by evaluating reliability diagrams (Hamill 1997), which compare forecast probabilities to their observed frequencies, with close

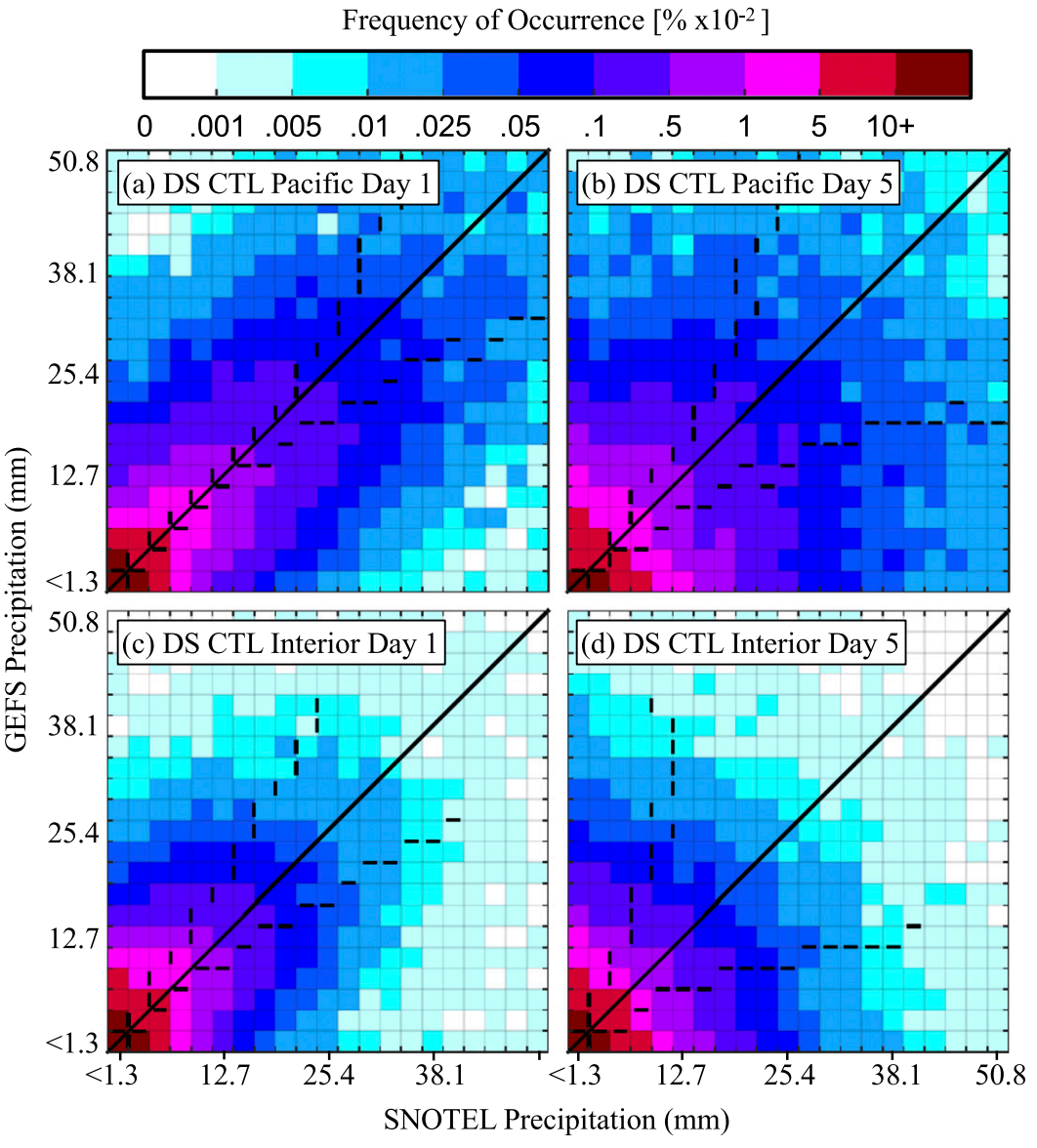

FIG. 10. As in Fig. 7, but for the downscaled GEFS.

correspondence indicating a reliable ensemble forecast system (Toth et al. 2003). In the Pacific ranges, reliability diagrams for day 1 PQPFs exhibit a slope much less than 1, indicating that the GEFS is strongly overconfident (i.e., underdispersive) for short-range forecasting (Fig. 14a). Events occur more frequently than predicted when the GEFS produces a low-probability forecast and less frequently than predicted when producing a medium- to high-probability forecast. Similar but somewhat lower reliability occurs in the interior ranges (Fig. 14b). Reliability over the Pacific ranges improves through day 5 for low-probability forecasts (i.e., <50%), but exhibits similar overconfidence for high-probability forecasts (cf. Figs. 14a,c). The improvement over the interior ranges by day 5 is smaller, and medium- to high-probability forecasts remain strongly overconfident (cf. Figs. 14b,d).

Ideally, a probabilistic system exhibits both reliability and *sharpness* (the relative magnitude of the ensemble spread), with an unreliable yet sharp system being undesirable (Toth et al. 2003). However, in addition to overconfidence, day 1 GEFS PQPFs are relatively sharp and frequently produce extreme low (0%) and high forecast (100%) probabilities in both the Pacific and interior ranges (Figs. 14a,b, inset). Sharpness decreases by forecast day 5 across the western United States as extreme low and high forecast probabilities are issued less frequently (Figs. 14c,d, inset).

The BSS (Brier 1950) indicates how a probabilistic system performs relative to the climatological event

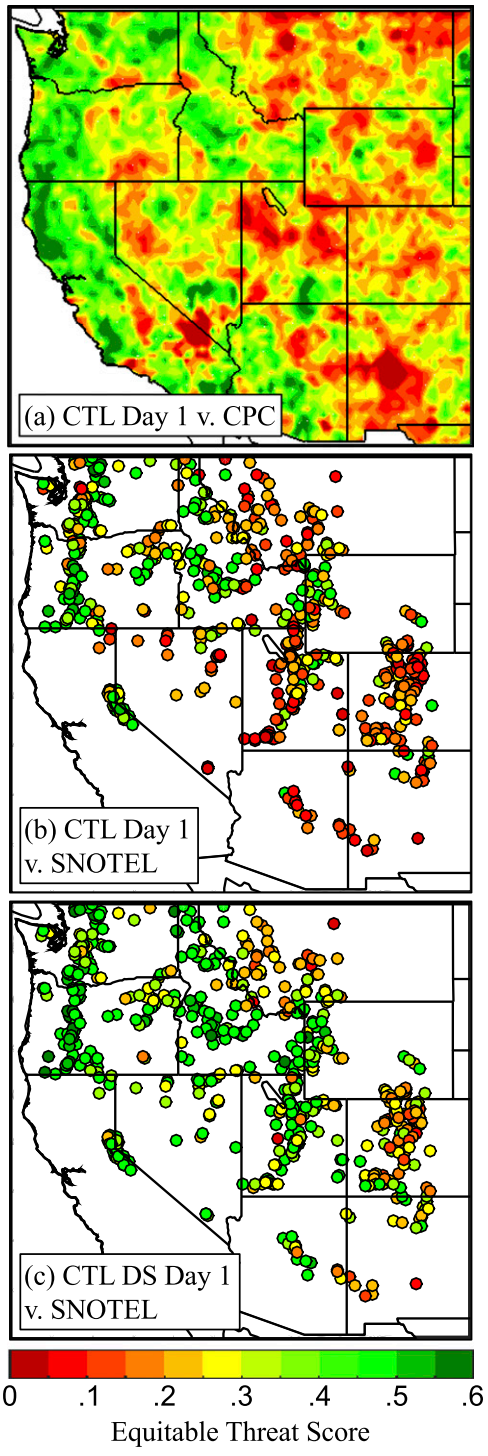

FIG. 11. ETSs for upper-quartile daily precipitation events. (a) GEFS day 1 (12–36 h) CTL forecasts relative to the CPC analysis. (b) GEFS day 1 CTL relative to SNOTEL observations. (c) GEFS day 1 downscaled CTL relative to SNOTEL observations.

frequencies obtained from the sample climatology (Toth et al. 2003). A perfect BSS is 1.0, a BSS of 0.0 indicates no skill over climatology, and a negative BSS indicates skill lower than climatology. The GEFS BSSs are positive in the Pacific ranges at days 1 and 5, indicating some skill relative to climatology, although the skill by day 5 is minimal (BSS = 0.17; Figs. 14a,c). BSSs over the interior ranges are smaller and only slightly positive on day 5, indicating that GEFS PQPFs exhibit minimal skill in comparison to climatological probabilities (Figs. 14b,d).

Rank histograms illustrate where observations fall within the ensemble distribution when sorted from low to high values (Hamill 2001). Typically, the desired result is that observations are equally likely to occur between any two ensemble members. While GEFS PQPFs in the Pacific ranges generally produce a larger ensemble spread and capture 9% (14%) more of the upper-quartile events at day 1 (day 5) than in the interior ranges, we present rank histograms for all SNOTEL stations since the underlying themes of the results are generally similar.

Consistent with the aforementioned problems predicting larger events, the day 1 ensemble spread captures only 18% of the upper-quartile events, with precipitation amounts during ~80% of those events exceeding the wettest ensemble member (Fig. 15a). Upper-quartile events with less precipitation than predicted by the driest ensemble member are relatively rare (3%).

Relatively large ensemble spreads are infrequent at day 1 (Fig. 15a, inset), which reflects the sharp and underdispersive nature of the GEFS for short-range forecasting. Larger ensemble spread sizes occur more frequently at longer lead times, such as day 5 (Fig. 15b, inset), allowing the spread to capture 29% of events. However, precipitation during ~70% of the events still exceeds the wettest ensemble member.

Downscaled GEFS day 1 and day 5 PQPFs share similar reliability diagram properties compared with the undownscaled GEFS (Fig. 14). While downscaling improves the reliability of lower forecast probabilities, higher forecast probabilities are less reliable. Downscaling inherently yields PQPFs that are less sharp as a result of the enhancement of GEFS QPFs at high-elevation SNOTEL stations (Fig. 14, inset). Downscaling worsens the BSSs over the interior ranges at day 1 and yields a relatively small improvement at day 5 and in the Pacific ranges (Fig. 14).

Downscaling reduces sharpness and improves the portion of upper-quartile events captured by the day 1 ensemble spread from 18% to 38% (cf. Fig. 15a,c). About 10% of events are overpredicted by all downscaled ensemble members at day 1, while ~50% are

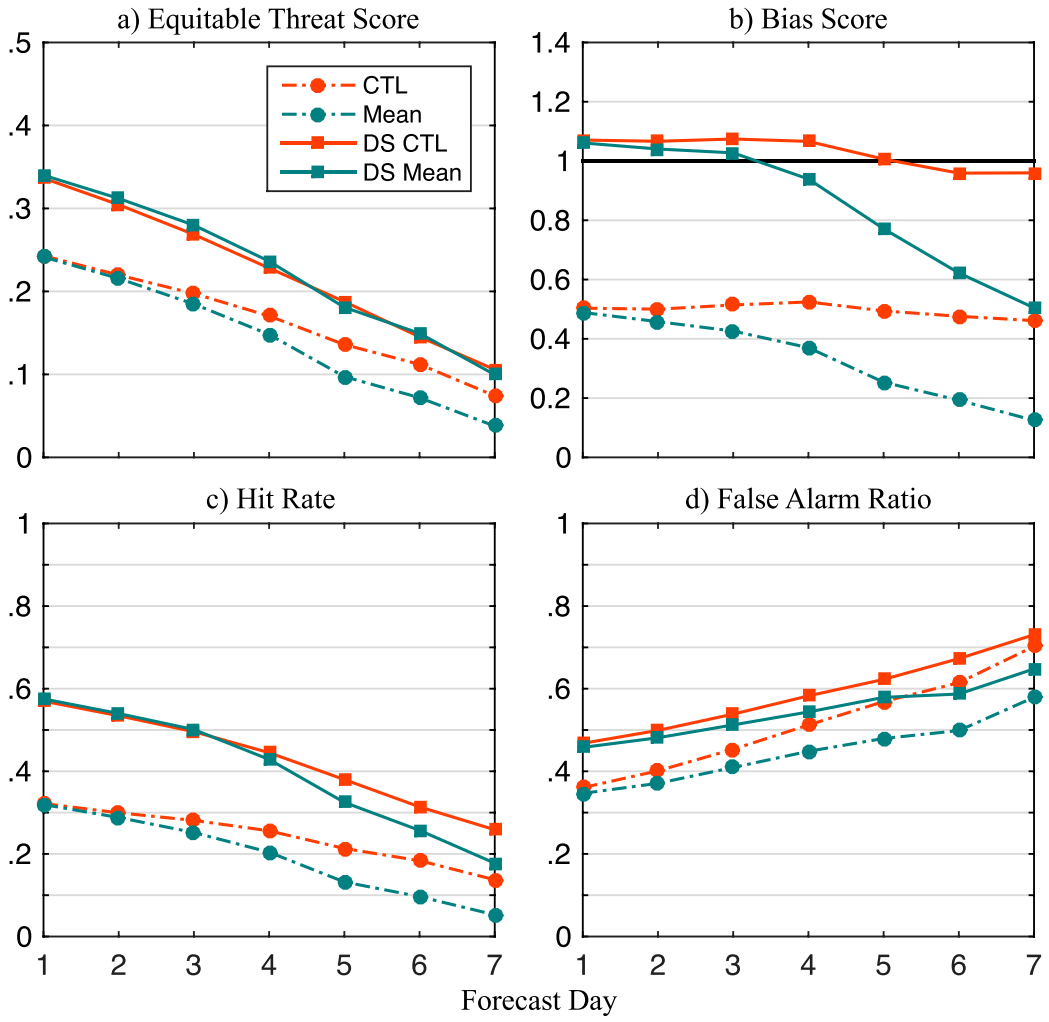

FIG. 12. Statistical measures of GEFS CTL (red dash–dot line/circle), GEFS ensemble mean (mean, teal dash–dot line/circle), downscaled GEFS control (DS CTL, red line/square), and downscaled GEFS ensemble mean (DS mean, teal line/square) forecasts of upper-quartile precipitation events at SNOTEL stations with increasing forecast lead time: (a) ETS, (b) bias score, (c) hit rate, and (d) false alarm ratio.

underpredicted (Figs. 15c). The downscaled ensemble spread is expectedly larger at day 5 such that 56% of upper-quartile events are captured, while ~45% remain underpredicted (Fig. 15d).

4. Conclusions

We have evaluated three cool seasons (October–March) of reforecasts and forecasts produced by the operational GEFS over the western United States using the CPC analysis to identify broad regional biases and SNOTEL observations for gauge-based validation in upper-elevation regions. Validation against the CPC precipitation analysis shows that the GEFS control (as well as individual members) generally produces too little precipitation over and upstream of topographic barriers and too much precipitation in downstream valleys and basins. Relative to SNOTEL observations, which are preferentially located in relatively wet upper-elevation regions, the GEFS control (and other individual members) has a pronounced dry bias at most locations. This dry bias reflects the infrequent production of larger 24-h precipitation events [i.e., ≳22.9 mm (10.2 mm) at stations in the Pacific (interior) ranges] relative to observations. Bivariate histograms show that at all but the smallest thresholds, observed events are more likely than not to be underforecast, with a greater likelihood in the interior.

For traditional performance measures [e.g., equitable threat score (ETS), hit rate, bias score, and false alarm ratio], the performance of the GEFS control (and other individual members) for upper-quartile precipitation events is highest in the Pacific ranges and generally

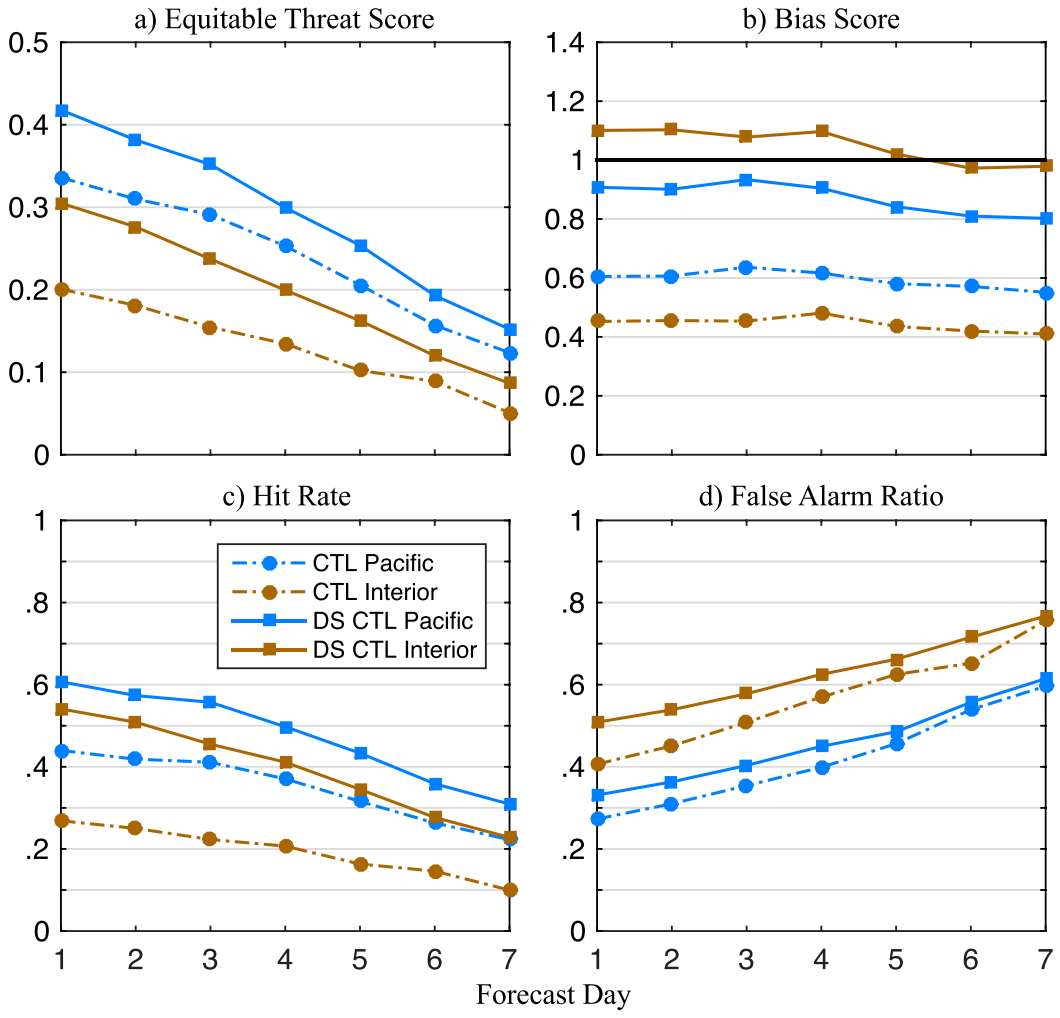

FIG. 13. Regional statistical measures of GEFS CTL (dash–dot line/circle) and downscaled GEFS control (DS CTL, line/square) forecasts for upper-quartile precipitation events at Pacific (blue) and interior (brown) SNOTEL stations with increasing forecast lead time. Shown are the (a) ETS, (b) bias score, (c) hit rate, and (d) false alarm ratio.

degrades toward the interior with considerable spatial variability. Based on ETS, a day 5 forecast over the Pacific ranges is as skillful as a day 1 forecast over the interior ranges. Hit rates and false alarm ratios are best at day 1, when the GEFS control upper-quartile-event hit rate is 0.44 (0.27) in the Pacific (interior) ranges, and the false alarm ratio is 0.27 (0.41).

Probabilistic verification statistics reflect both the underprediction biases inherent in the GEFS control (and individual members), as well as the unreliable (or overconfident) and underdispersive nature of the GEFS. Observed upper-quartile precipitation events at SNOTEL stations exceed the wettest member of the GEFS ensemble at day 1 (day 5) ~80% (~70%) of the time. At day 1, PQPFs for upper-quartile events are strongly overconfident, with low-probability (high probability) forecasts associated with a higher (lower) frequency of observed events. Reliability improves with increasing forecast lead time, but high-probability forecast overconfidence is still evident at day 5. Forecasters should be aware that although the GEFS has a low frequency bias for larger events, a high PQPF for a larger event is likely an overestimate of the actual event probability. Day 1 and day 5 PQPFs for upper-quartile events in the Pacific ranges are more skillful than using climatological probabilities (BSS = 0.28 and 0.17, respectively), but over the interior ranges, such PQPFs exhibit minimal improvements over climatological probabilities (BSS = 0.14 and 0.06, respectively).

In an attempt to improve GEFS QPFs, we produced statistically downscaled forecasts derived from high-resolution climatological precipitation analyses produced

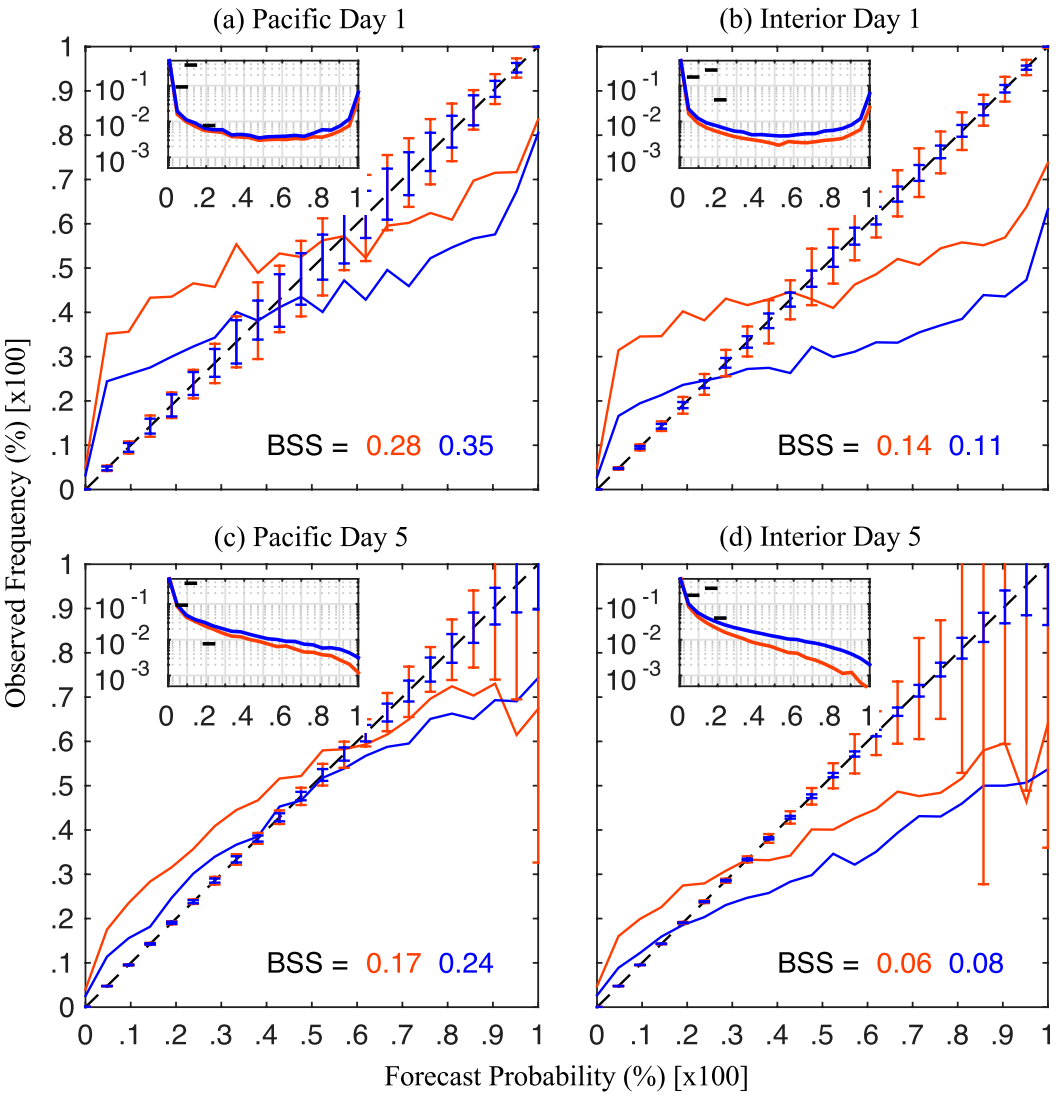

FIG. 14. Reliability diagrams for GEFS day 1 (12–36 h) (red) and downscaled GEFS day 1 (blue) forecasts of upper-quartile events at (a) Pacific and (b) interior range SNOTEL stations. (c),(d) As in (a),(b), but for day 5 (108–132 h). BSS is annotated. Inset histograms indicate the relative frequency of forecast probabilities for GEFS (red) and downscaled GEFS (blue) forecasts, as well as the climatological event frequencies (black lines). Whiskers represent 5% and 95% confidence bars.

by the PRISM Climate Group at Oregon State University. Such downscaling generally resolves dry biases at SNOTEL locations, as well as the tendency for most events to be underforecast. Downscaling also improves ETSs, hit rates, and bias scores. For example, the downscaled GEFS control ETS for upper-quartile events at day 4 is roughly equivalent to that of the undownscaled GEFS control at day 1. However, upper-quartile-event false alarm ratios at day 1 are worsened to 0.33 (0.51) in the Pacific (interior) ranges. Thus, forecasters should recognize that while downscaling improves ETSs, hit rates, and bias scores for upper-quartile events, it also increases false alarms.

For PQPFs, downscaling worsens reliability by exacerbating the overconfidence of high-probability forecasts. However, at day 1 (day 5), 38% (56%) of upper-quartile events are captured by the downscaled ensemble spread, which is an improvement over the undownscaled GEFS. Nevertheless, most missed events are underforecast by the wettest ensemble member (rather than overforecast by the driest member). Downscaled PQPFs in the Pacific ranges have slightly improved BSSs, while BSSs in the interior ranges change minimally.

These findings indicate significant limitations in GEFS QPF and PQPF over the western United States as a result of insufficient resolution and underdispersion.

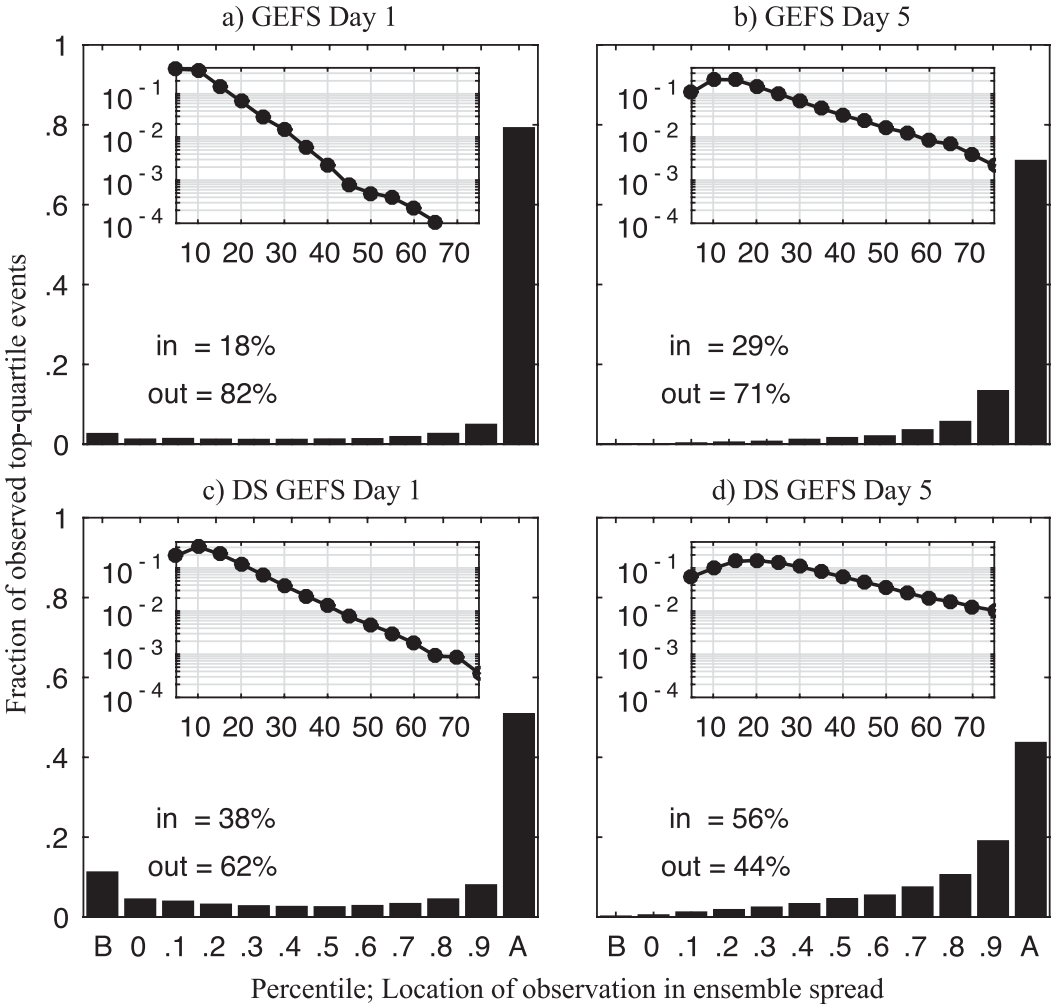

FIG. 15. Rank histograms for (a) GEFS day 1 (12–36 h), (b) downscaled GEFS day 1, (c) GEFS day 5 (108–132 h), and (d) downscaled GEFS day 5 forecasts of upper-quartile precipitation events observed at all SNOTEL stations. Letters A and B indicate above and below the ensemble spread, respectively. Insert histograms indicate the frequency of ensemble spread size in 5-mm bins. Annotations of "in" and "out" reflect percentages of the observations occurring inside and outside of the ensemble spread, respectively.

These limitations are especially acute for larger events over the finescale topography of the western interior. Efforts to improve GEFS precipitation forecasts through climatology-based downscaling yield some improvements, but also an increase in false alarms, especially over the interior. The extent to which these results are exacerbated by the relatively low-resolution 1.0° grid is unclear and perhaps some improvement would occur with a higher-resolution output grid, such as the 0.5° latitude–longitude grid that is now available. However, even at native grid spacing (~33 km), finescale orographic effects remain unresolved. Western U.S. forecasters should be aware of the capabilities and limitations of the GEFS and downscaled GEFS identified herein. Future work should examine the performance of alternative downscaling and ensemble calibration approaches as these may offer a pathway to improved forecasts by better accounting for model bias, regime-dependent variations in orographic enhancement, and probabilistic properties of the ensemble.

Acknowledgments. We thank the NOAA/Earth System Research Laboratory Physical Sciences Division for providing CPC precipitation data, the NRCS for providing SNOTEL data, the PRISM Climate Group at Oregon State University for providing PRISM data, NOAA/NCEP for providing GEFS forecasts and reforecasts, and the University of Utah Center for High Performance Computing for computer-support services. This work is an outgrowth of master of science research

conducted by WRL at the University of Utah, and we thank committee members Court Strong and Larry Dunn for their input. This article is based on research supported by the NOAA/National Weather Service CSTAR Program through Grant NA13NWS4680003. 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

Avanzi, F., C. D. Michele, A. Ghezzi, C. Jommi, and M. Pepe, 2014: A processing–modeling routine to use SNOTEL hourly data in snowpack dynamic models. *Adv. Water Resour.*, **73**, 16–29, doi:10.1016/j.advwatres.2014.06.011.

Baxter, M. A., G. M. Lackmann, K. M. Mahoney, T. E. Workoff, and T. M. Hamill, 2014: Verification of quantitative precipitation reforecasts over the southeastern United States. *Wea. Forecasting*, **29**, 1199–1207, doi:10.1175/WAF-D-14-00055.1.

Black, A. W., and T. L. Mote, 2015: Characteristics of winter-precipitation-related transportation fatalities in the United States. *Wea. Climate Soc.*, **7**, 133–144, doi:10.1175/WCAS-D-14-00011.1.

Bontron, G., and C. Obled, 2005: A probabilistic adaptation of meteorological model outputs to hydrological forecasting. *Houille Blanche*, **1**, 23–28, doi:10.1051/lhb:200501002.

Brier, G. W., 1950: The statistical theory of turbulence and the problem of diffusion in the atmosphere. *J. Meteor.*, **7**, 283–290, doi:10.1175/1520-0469(1950)007<0283:TSTOTA>2.0.CO;2.

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

Charles, M. E., and B. A. Colle, 2009: Verification of extratropical cyclones with the NCEP operational models. Part I: Analysis errors and short-term NAM and GFS forecasts. *Wea. Forecasting*, **24**, 1173–1190, doi:10.1175/2009WAF2222169.1.

Chen, M., W. Shi, P. Xie, V. B. S. Silva, V. E. Kousky, R. W. Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.*, **113**, D04110, doi:10.1029/2007JD009132.

Cohen, J., 1996: Snowstorms. *Encyclopedia of Weather and Climate*, S. H. Schneider, Ed., Vol. 2, Oxford University Press, 700–703.

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, doi: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, doi:10.1002/joc.1688.

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, **13**, 1132–1147, doi:10.1175/1520-0434(1998)013<1132:CPQPFB>2.0.CO;2.

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, doi:10.1002/hyp.5806.

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138**, 190–202, doi:10.1175/2009MWR3046.1.

Gutmann, E. D., R. M. Rasmussen, C. Liu, K. Ikeda, D. J. Gochis, M. P. Clark, J. Dudhia, and G. Thompson, 2012: A comparison of statistical and dynamic downscaling of winter precipitation over complex terrain. *J. Climate*, **25**, 262–281, doi:10.1175/2011JCLI4109.1.

Hamill, T. M., 1997: Reliability diagrams for multicategory probabilistic forecasts. *Wea. Forecasting*, **12**, 736–741, doi:10.1175/1520-0434(1997)012<0736:RDFMPF>2.0.CO;2.

——, 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, 2012: Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Mon. Wea. Rev.*, **140**, 2232–2252, doi:10.1175/MWR-D-11-00220.1.

——, and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, doi:10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

——, and ——, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724, doi:10.1175/1520-0493(1998)126<0711:EOEREP>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, doi:10.1256/qj.06.25.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, doi:10.1175/MWR3237.1.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, doi:10.1175/2007MWR2411.1.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, doi:10.1175/BAMS-D-12-00014.1.

——, M. Scheuerer, and G. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143**, 3300–3309, doi:10.1175/MWR-D-15-0004.1.

Hatchett, B. J., S. Burak, J. J. Rutz, N. S. Oakley, E. H. Bair, and M. Kaplan, 2017: Avalanche fatalities during atmospheric river events in the United States. *J. Hydrometeor.*, doi:10.1175/JHM-D-16-0219.1, in press.

Higgins, R. W., W. Shi, E. Yarosh, and R. Joyce, 2000: Improved United States precipitation quality control system and analysis. NCEP/Climate Prediction Center Atlas 7, NCEP/CPC. [Available online at http://www.cpc.ncep.noaa.gov/products/outreach/research_papers/ncep_cpc_atlas/7/.]

Hou, D., and Coauthors, 2014: Climatology-Calibrated Precipitation Analysis at fine scales: Statistical adjustment of stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557, doi:10.1175/JHM-D-11-0140.1.

——, Y. Zhu, X. Zhou, R. Wobus, J. Peng, Y. Luo, and B. Cui, 2015: The 2015 upgrade of NCEP's Global Ensemble Forecast System (GEFS). *27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction*, Chicago, IL, Amer. Meteor. Soc., 2A.6. [Available online at https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273406.html.]

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

Junker, N. W., J. E. Hoke, B. E. Sullivan, K. F. Brill, and F. J. Hughes, 1992: Seasonal geographic variations in quantitative precipitation prediction by NMC's nested-grid model and medium-range forecast model. *Wea. Forecasting*, **7**, 410–429, doi:10.1175/1520-0434(1992)007<0410:SAGVIQ>2.0.CO;2.

Kunz, M., and C. Kottmeier, 2006: Orographic enhancement of precipitation over low mountain ranges. Part II: Simulations of heavy precipitation events over southwest Germany. *J. Appl. Meteor. Climatol.*, **45**, 1041–1055, doi:10.1175/JAM2390.1.

Marsigli, C., A. Montani, F. Nerozzi, T. Paccagnella, S. Tibaldi, F. Molteni, and R. Buizza, 2001: A strategy for high-resolution ensemble prediction. Part II: Limited-area experiments in four Alpine flood events. *Quart. J. Roy. Meteor. Soc.*, **127**, 2095–2115, doi:10.1002/qj.49712757613.

Mason, I. B., 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.

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **129**, 638–663, doi:10.1175/1520-0493(2001)129<0638:QPFOTU>2.0.CO;2.

Neiman, P. J., L. J. Schick, F. M. Ralph, M. Hughes, and G. A. Wick, 2011: Flooding in western Washington: The connection to atmospheric rivers. *J. Hydrometeor.*, **12**, 1337–1358, doi:10.1175/2011JHM1358.1.

NOAA, 2015: Technical implementation notice 15-43. NOAA/National Weather Service. [Available online at http://www.nws.noaa.gov/os/notification/tin15-43gefs.htm.]

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, doi:10.1002/joc.4504.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi:10.1175/MWR2906.1.

Ralph, F. M., P. J. Neiman, G. A. Wick, S. I. Gutman, M. D. Dettinger, D. R. Cayan, and A. B. White, 2006: Flooding on California's Russian River: Role of atmospheric rivers. *Geophys. Res. Lett.*, **33**, L13801, doi:10.1029/2006GL026689.

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, doi:10.1175/BAMS-D-11-00052.1.

Rutledge, G. K., J. Alpert, and W. Ebisuzaki, 2006: NOMADS: A climate and weather model archive at the National Oceanic and Atmospheric Administration. *Bull. Amer. Meteor. Soc.*, **87**, 327–341, doi:10.1175/BAMS-87-3-327.

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

——, ——, 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, doi: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, doi: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, doi:10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, doi:10.1175/MWR-D-15-0061.1.

Schirmer, M., and B. Jamieson, 2015: Verification of analyzed and forecasted winter precipitation in complex terrain. *Cryosphere*, **9**, 587–601, doi:10.5194/tc-9-587-2015.

Schmeits, M. J., and K. J. Kok, 2010: A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Mon. Wea. Rev.*, **138**, 4199–4211, doi:10.1175/2010MWR3285.1.

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, doi:10.1029/1999WR900090.

——, ——, and A. Frei, 2001: Characteristics of larger snowfall events in the montane western United States as examined using snowpack telemetry (SNOTEL) data. *Water Resour. Res.*, **37**, 675–688, doi:10.1029/2000WR900307.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, doi:10.1175/MWR3441.1.

Smith, B. L., S. E. Yuter, P. J. Neiman, and D. E. Kingsmill, 2010: Water vapor fluxes and orographic precipitation over northern California associated with a landfalling atmospheric river. *Mon. Wea. Rev.*, **138**, 74–100, doi:10.1175/2009MWR2939.1.

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

——, 2004: One hundred inches in one hundred hours: The complex evolution of an intermountain winter storm cycle. *Bull. Amer. Meteor. Soc.*, **85**, 16–20, doi:10.1175/BAMS-85-1-16.

——, 2014: *Secrets of the Greatest Snow on Earth.* Utah State University Press, 244 pp.

Stensrud, D. J., H. E. Brooks, J. Du, S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433–446, doi:10.1175/1520-0493(1999)127<0433:UEFSRF>2.0.CO;2.

Stewart, R. E., and Coauthors, 1995: Winter storms over Canada. *Atmos.–Ocean*, **33**, 223–247, doi:10.1080/07055900.1995.9649533.

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. Stephenson, Eds., John Wiley and Sons, 137–163.

Tremper, B., 2008: *Staying Alive in Avalanche Terrain.* Mountaineers Books, 318 pp.

U.S. Department of the Interior, 2012: Flood of January 1997 in the Truckee River basin, western Nevada. USGS Fact Sheet FS-123-97, 2 pp. [Available online at http://pubs.usgs.gov/fs/1997/0123/report.pdf.]

Van Haren, R., R. J. Haarsma, G. J. Van Oldenborgh, and W. Hazeleger, 2015: Resolution dependence of European precipitation in a state-of-the-art atmospheric general circulation model. *J. Climate*, **28**, 5134–5149, doi:10.1175/JCLI-D-14-00279.1.

Wang, X., D. Parrish, D. Kleist, and J. Whitaker, 2013: GSI 3DVar-based ensemble–variational hybrid data assimilation for NCEP

Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.*, **141**, 4098–4117, doi:10.1175/MWR-D-12-00141.1.

Wilby, R. L., T. M. L. Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks, 1998: Statistical downscaling of general circulation model output: A comparison of methods. *Water Resour. Res.*, **34**, 2995–3008, doi:10.1029/98WR02577.

Wilks, D. S., 2006a: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Appl.*, **13**, 243–256, doi:10.1017/S1350482706002192.

——, 2006b: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.*, **107**, 4429, doi:10.1029/2001JD000659.

——, L. Leung, V. Sridhar, and D. Lettenmaier, 2004: Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, **62**, 189–216, doi:10.1023/B:CLIM.0000013685.99609.9e.

Xie, P., A. Yatagai, M. Chen, T. Hayasaka, Y. Fukushima, C. Liu, and S. Yang, 2007: A gauge-based analysis of daily precipitation over East Asia. *J. Hydrometeor.*, **8**, 607–626, doi:10.1175/JHM583.1.

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, doi:10.1175/1520-0426(1998)015<0054:AONSNP>2.0.CO;2.

Yuan, H., S. L. Mullen, X. Gao, S. Sorooshian, J. Du, and H. H. Juang, 2005: Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system. *Mon. Wea. Rev.*, **133**, 279–294, doi:10.1175/MWR-2858.1.