Validation of Cool-Season Snowfall Forecasts at a High-Elevation Site in Utah's Little Cottonwood Canyon

MICHAEL D. PLETCHER,[a] PETER G. VEALS,[a] MICHAEL E. WESSLER,[b] DAVID CHURCH,[b] KIRSTIN HARNOS,[c] JAMES CORREIA JR.,[c] RANDY J. CHASE,[d] AND W. JAMES STEENBURGH[a]

[a] *Department of Atmospheric Sciences, University of Utah, Salt Lake City, Utah*
[b] *National Weather Service, Salt Lake City, Utah*
[c] *NOAA/NWS/Weather Prediction Center, College Park, Maryland*
[d] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*

(Manuscript received 19 October 2023, in final form 19 April 2024, accepted 2 July 2024)

ABSTRACT: Producing a quantitative snowfall forecast (QSF) typically requires a model quantitative precipitation forecast (QPF) and snow-to-liquid ratio (SLR) estimate. QPF and SLR can vary significantly in space and time over complex terrain, necessitating fine-scale or point-specific forecasts of each component. Little Cottonwood Canyon (LCC) in Utah's Wasatch Range frequently experiences high-impact winter storms and avalanche closures that result in substantial transportation and economic disruptions, making it an excellent testbed for evaluating snowfall forecasts. In this study, we validate QPFs, SLR forecasts, and QSFs produced by or derived from the Global Forecast System (GFS) and High-Resolution Rapid Refresh (HRRR) using liquid precipitation equivalent (LPE) and snowfall observations collected during the 2019/20–2022/23 cool seasons at the Alta–Collins snow-study site (2945 m MSL) in upper LCC. The 12-h QPFs produced by the GFS and HRRR underpredict the total LPE during the four cool seasons by 33% and 29%, respectively, and underpredict 50th, 75th, and 90th percentile event frequencies. Current operational SLR methods exhibit mean absolute errors of 4.5–7.7. In contrast, a locally trained random forest algorithm reduces SLR mean absolute errors to 3.7. Despite the random forest producing more accurate SLR forecasts, QSFs derived from operational SLR methods produce higher critical success indices since they exhibit positive SLR biases that offset negative QPF biases. These results indicate an overall underprediction of LPE by operational models in upper LCC and illustrate the need to identify sources of QSF bias to enhance QSF performance.

SIGNIFICANCE STATEMENT: Winter storms in mountainous terrain can disrupt transportation and threaten life and property due to road snow and avalanche hazards. Snow-to-liquid ratio (SLR) is an important variable for snowfall and avalanche forecasts. Using high-quality historical snowfall observations and atmospheric analyses, we developed a machine learning technique for predicting SLR at a high mountain site in Utah's Little Cottonwood Canyon that is prone to closure due to winter storms. This technique produces improved SLR forecasts for use by weather forecasters and snow-safety personnel. We also show that current operational models and SLR techniques underforecast liquid precipitation amounts and overforecast SLRs, respectively, which has implications for future model development.

KEYWORDS: Snow; Winter/cool season; Orographic effects; Forecasting

1. Introduction

Quantitative snowfall forecasts (QSFs) over complex terrain pose challenges for operational forecasters due to the localized nature of snow accumulation patterns and uncertainties in the specification of snow-to-liquid ratio (SLR) (e.g., Alcott and Steenburgh 2010; Mott et al. 2014; Gerber et al. 2018, 2019). Despite a shift to a greater fraction of precipitation falling as rain due to climate change (Knowles et al. 2006; Gillies et al. 2012), cool-season (November–April) precipitation over the western contiguous United States (CONUS) still falls predominantly as snow at upper elevations, and many low-elevation regions experience episodic winter storms that can disrupt transportation and commerce (e.g., Ferber et al. 1993; Daly et al. 1994; Andretta and Hazen 1998; Serreze et al. 1999; Alcott et al. 2012). High-impact, cool-season precipitation events can be associated with diverse synoptic patterns that require accurate precipitation type identification and snowfall forecasts to reduce travel disruptions, loss of life, and damage to property and infrastructure (Lackmann and Gyakum 1999; Steenburgh 2003; Ralph et al. 2006; Rutz et al. 2014; Seeherman and Liu 2015; Wasserstein and Steenburgh 2024). State Route 210 (SR-210) in Utah's Little Cottonwood Canyon (LCC) is one example of a mountain transportation route that can temporarily close due to heavy snowfall, which can gridlock transportation and force avalanche mitigation activities (Nalli and Mckee 2018). During busy travel periods, the number of vehicles traveling in LCC can exceed 10 000 per day, and road closures due to avalanche hazards or mitigation can result in financial losses to local ski resorts of more than $3 million per day (amount adjusted to 2023 dollars; Blattenberger and Fowles 1995).

Contemporary operational numerical forecast systems do not explicitly predict QSF. Instead, after determining snow is the predominant precipitation type, which can be done with

Corresponding author: Michael D. Pletcher, michael.pletcher@utah.edu

DOI: 10.1175/WAF-D-23-0176.1

different approaches (e.g., Bourgouin 2000; Reeves et al. 2016; Benjamin et al. 2016; Birk et al. 2021), an SLR is typically applied to the model quantitative precipitation forecast (QPF) or a downscaled or statistically adjusted QPF (e.g., Lewis et al. 2017; Hamill and Scheuerer 2018) to generate a QSF. A deterministic multimodel QSF or probabilistic QSF (PQSF) can be generated using ensemble forecast systems or blends of forecast systems (Craven et al. 2020).

Thus, the model QPF is one source of error when generating QSFs. Over complex terrain, underprediction of QPF is a common bias in lower-resolution modeling systems (e.g., Lewis et al. 2017). Smaller horizontal grid spacings enable better resolution of topographic features, generally resulting in increased skill and smaller biases, especially over narrow terrain features such as those found in the Great Basin (e.g., Mass et al. 2002; Hart et al. 2005; Gowan et al. 2018). Postprocessing techniques such as downscaling with climatological analyses (e.g., Daly et al. 2008; Lewis et al. 2017; Riley et al. 2021; Stovern et al. 2023), quantile mapping (e.g., Hamill et al. 2023), bias correction (e.g., Velasquez et al. 2020), and machine learning (e.g., Sha et al. 2020; Espeholt et al. 2022) have been used to produce improved QPF and probabilistic QPF (PQPF) in regions of complex terrain. Such approaches may reduce bias but do not eliminate errors related to initial conditions or model uncertainties.

SLR uncertainty is another source of QSF error. Historically, and even on many meteorological websites, SLR is sometimes assumed to follow the 10:1 rule [e.g., 10 mm of QPF produces 10 cm of QSF (Potter 1965; Judson and Doesken 2000)]. In reality, SLR exhibits considerable geographic and temporal variability with values ranging from 3:1 to 100:1 in rare events (Judson and Doesken 2000; Roebber et al. 2003; Baxter et al. 2005; Alcott and Steenburgh 2010). SLR is influenced by ice crystal habit which varies based on the degree of aggregation, riming, sublimation, melting of peripheral ice crystal branches at the surface or aloft, or crystal fragmentation by strong winds; SLR is also affected by melting at the surface or aloft, precipitation-type changes, and surface-snow compaction (Pomeroy and Brun 2001; Dubé 2003; Roebber et al. 2003; Baxter et al. 2005; Cobb and Waldstreicher 2005; Byun et al. 2008; Alcott and Steenburgh 2010; Milbrandt et al. 2012). Due to the varying weather conditions that influence SLR, it can be a challenge to predict storm mean SLR. Changes in SLR during storms are also difficult to predict but are important for diagnosing avalanche hazards (e.g., Mueller 2001; Schweizer et al. 2003; Schweizer and Reuter 2015). SLR can also vary significantly with elevation and distance over relatively small geographic regions. A variety of SLR algorithms have been produced using physical reasoning or statistical approaches (e.g., Dubé 2003; Roebber et al. 2003; Cobb and Waldstreicher 2005; Alcott and Steenburgh 2010; Milbrandt et al. 2012; Hoopes et al. 2023). Despite advancements in SLR algorithms and understanding, SLR variability remains a significant source of uncertainty for QSF.

In this paper, we examine the performance of precipitation, SLR, and snowfall forecasts produced by or derived from the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) and High-Resolution Rapid Refresh (HRRR) for the Alta–Collins (CLN) snow-study site maintained by Alta Ski Area. Located in the upper reaches of a deeply incised canyon in the Wasatch Range of northern Utah (Fig. 1), CLN observes frequent heavy snowfall during the cool season, defined here as November–April, although snow can fall outside this period, with road snow and avalanche hazards posing major threats to travel, structures, and public safety. The mean annual snowfall at the National Weather Service (NWS) Cooperative Observer Program (COOP) site in the town of Alta (2655 m MSL) in upper LCC is 1164 cm [(458 in.; 1991–2020 climate normal (NCEI 2023)]. SR-210, which ascends LCC to the town of Alta, traverses 50 avalanche paths and has the highest uncontrolled avalanche hazard index of any major road in the world due to the high frequency of major avalanches and heavy vehicle traffic (Schaerer 1989; Nalli and Mckee 2018; Steenburgh 2023; Wasserstein and Steenburgh 2024). Avalanche mitigation activities or unmitigated avalanche threats close SR-210 an average of 10.8 days a year and more than 30 days in heavy snow years (Utah Department of Transportation 2022). In 2023, multiday closures of SR-210 stranded tourists and residents, leading to food shortages (Jag 2023). Both total snowfall amount and rate contribute to avalanche hazards, as do changes in SLR (Mueller 2001; Schweizer et al. 2003; Schweizer and Reuter 2015). Rainfall can also be important for avalanches, but most of the cool-season precipitation in avalanche starting zones in LCC falls as snow.

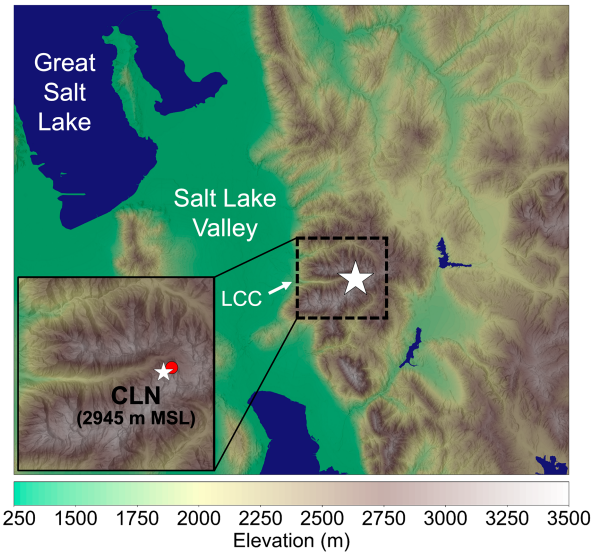

FIG. 1. Topography and geographic features of northern Utah. Zoomed-in topography of LCC and locations of CLN (40.5763°N, 111.6383°W) and Alta GFS and HRRR BUFKIT profiles (40.58°N, 111.63°W) indicated by the white star and red dot, respectively, in the inset map at the lower left.

The remainder of this paper is organized as follows. In section 2, we detail our data and methods including a description of the CLN observations, SLR forecast methods, and validation techniques. Section 3 reveals the seasonal QPF biases exhibited by the GFS and HRRR and their QPF performance

for individual precipitation events. Section 4 highlights discrepancies in forecast SLR between the GFS and HRRR and the biases exhibited by each SLR method. Section 5 focuses attention on the seasonal QSFs derived from each SLR method, QSF performance for individual snowfall events, and QSF biases exhibited relative to LPE and snowfall event sizes. Finally, section 6 summarizes the key conclusions, discusses the importance of validating QPF and SLR errors for improving QSFs, and provides suggestions for future work.

2. Data and methods

a. CLN observations

This study uses 12-h snowfall and liquid precipitation equivalent (LPE) measurements collected at 0400 and 1600 local standard time (LST) by snow-safety professionals at CLN during the 2004/05–2022/23 cool seasons (Wasserstein and Steenburgh 2023). The first 15 cool seasons (2004/05–2018/19) were used to train a new random forest (RF; Breiman 2001) SLR algorithm described in section 2b, whereas the last 4 cool seasons (2019/20–2022/23) were used to validate QPFs, SLR forecasts, and QSFs produced by or derived from the GFS and HRRR using the RF or other SLR methods. The focus on the last four cool seasons for validation reflects the upgrade to the GFS finite volume cubed-sphere dynamical core (FV3) in June 2019 (NCEP 2019). CLN is located at Alta Ski Area in upper LCC (2945 m MSL; Fig. 1) in a small clearing surrounded by evergreen trees and below ridgelines, which reduces wind influences (Alcott and Steenburgh 2010, see their Fig. 1). Snowfall observations were collected from a white snowboard placed atop the existing snowpack that was wiped clean after each measurement. LPE was obtained with a Snowmetrics sampling tube and scale. The minimum observed thresholds for LPE and snowfall measurements were 0.254 mm (0.01 in.) and 1.27 cm (0.5 in.) and were rounded to the nearest 0.5 and 0.01 in., respectively. These measurements were converted to metric units for this study and are used to evaluate the performance of precipitation and snowfall forecasts. SLR was based on the ratio of snowfall to LPE. For SLR calculations, however, we only used observations with at least 2.79 mm (0.11 in.) of LPE and 5.09 cm (2 in.) of snowfall. These thresholds have been used in other studies and reduce SLR errors arising from rounding and measurement inaccuracies (Judson and Doesken 2000; Roebber et al. 2003, 2007; Baxter et al. 2005; Alcott and Steenburgh 2010). For consistency, these thresholds are used to train the RF algorithm and to evaluate the performance of SLR forecasts produced by each method. In the present climate, nearly all precipitation at CLN falls as snow during the cool season, so rain on snow is rare and likely does not affect results.

b. Random forest SLR algorithm

Using data from the North American Regional Reanalysis (NARR; Mesinger et al. 2006) and eight cool seasons of 24-h snowfall observations, Alcott and Steenburgh (2010) developed an algorithm for predicting SLR at CLN based on stepwise multiple linear regression. For this study, we produced a new SLR algorithm using vertical profiles from the fifth major global reanalysis produced by European Centre for Medium-Range Weather Forecasts (ERA5; Hersbach et al. 2020), the 15-cool-season training period of SLR observations from CLN, and an RF machine learning algorithm. RFs are statistical algorithms that utilize an ensemble of decision trees. During the RF training, randomness is introduced by selecting random subsets of data and features for building each tree. RF predictions are then made deterministically by aggregating predictions from all trees in the ensemble. The RF training used the Python scikit-learn software package (Pedregosa et al. 2011) and a randomized 60/40 train–validate split for testing and development. We ultimately settled on a regressor with 100 trees, which appeared optimal based on the lowest mean absolute error (MAE) and highest coefficient of determination (R^2) scores we achieved during internal testing. Additional testing revealed that other hyperparameters (i.e., maximum tree depth and minimum number of samples required to split a node) produced optimal results when set to their default scikit-learn settings. The ERA5 reanalysis profile used for training was from the nearest grid point to CLN and contained data at 37 pressure levels from 1000 to 1 hPa. Although the ERA5 reanalysis is characterized by relatively low horizontal resolution compared to operational weather models, its extensive data record, assimilation of varying observation types, and consistent dynamical analysis make it a suitable training dataset.

The inputs fed into the RF algorithm were temperature (K) and wind (m s^{-1}) linearly interpolated to 0.5, 1, and 2 km above the elevation of CLN. We used temperature and wind as they exhibit the strongest (albeit modest) correlation with SLR (Roebber et al. 2003; Baxter et al. 2005; Cobb and Waldstreicher 2005; Alcott and Steenburgh 2010). We focus on 2 km above the elevation of CLN (4945 m MSL), which is typically just below 500 hPa. This is consistent with Alcott and Steenburgh (2010) who developed a stepwise multiple linear regression equation for predicting SLR at Alta and found that 15 of the 17 predictors were at or below 500 hPa. Internal testing using the ERA5 as the training dataset revealed that other potential SLR predictors (e.g., surface temperature, relative humidity, specific humidity, precipitable water, and geopotential height) did not contribute substantially to changes in predicted SLR. QPF is another potential predictor but was not used for this study to avoid complications arising from contrasting ERA5 and operational model precipitation biases.

During the 15-cool-season period used for training, there were 1204 SLR observations available for training (see section 2a) with 50th, 75th, and 90th percentile SLRs of 12.5, 16.8, and 21.2, respectively (Fig. 2c). During the four-cool-season validation period, the SLR distribution was shifted slightly higher with 50th, 75th, and 90th percentiles of 13.3, 17.9, and 23.2, respectively (Fig. 2f) although the 50th and 75th percentiles for both this and the training period are close to the 12.3 and 18 reported by Alcott and Steenburgh (2010) using 24-h observations from CLN from the 1998/99 to 2006/07 cool seasons (they did not report the 90th percentile). Overall, this suggests that the training period provides a reasonable long-

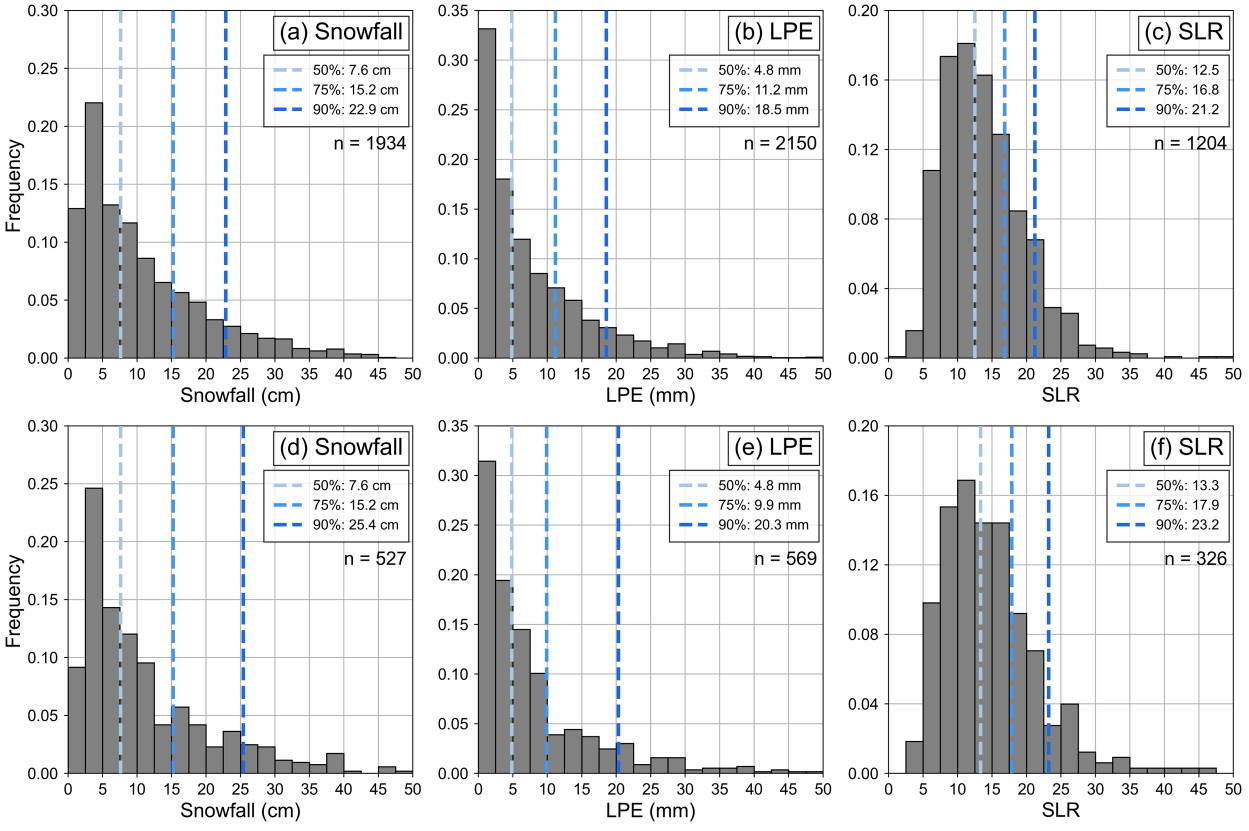

FIG. 2. One-dimensional histograms of CLN 12-h (a) snowfall, (b) LPE, and (c) SLR (RF training) collected between November 2004 and April 2019 with 50th, 75th, and 90th percentiles and number of events annotated. (d)–(f) As in (a)–(c), but for the validation dataset collected between November 2019 and April 2023.

term record suitable for machine learning SLR applications and that the snowfall and LPE frequency distributions for the validation period approximate those of a longer-term record.

c. *Operational SLR methods*

Several SLR prediction methods are used operationally by the NWS in the National Blend of Models (NBM), a CONUS-wide forecast system derived from direct and postprocessed model guidance (Craven et al. 2020). These methods are known as MaxTAloft, Cobb, Roebber, and 850–700-mb thickness (1 mb = 1 hPa), as described in NWS online training resources (The COMET Program 2023). Blends of these methods are then used depending on the forecast model (e.g., the SLR for the GFS is based on 33% weighting of the MaxTAloft, Cobb, and Roebber methods). The MaxTAloft method calculates SLR using a fifth-degree polynomial of the form

$$\begin{aligned}\mathrm{SLR}_{\mathrm{MaxTAloft}} &= 0.000\,004\,5 \times T_{\mathrm{Max}}^5 + 0.000\,443\,2 \times T_{\mathrm{Max}}^4 \\ &\quad + 0.013\,090\,3 \times T_{\mathrm{Max}}^3 + 0.058\,596\,8 \times T_{\mathrm{Max}}^2 \\ &\quad - 1.815\,080\,9 \times T_{\mathrm{Max}} + 5.980\,572\,2, \end{aligned} \quad (1)$$

where T_{max} is the maximum temperature in degrees Celsius between 2000 ft (610 m) AGL and 400 hPa. The polynomial is based on observations from Alaska and subjective adjustments (The COMET Program 2023). The Cobb method follows Cobb and Waldstreicher (2005) but has been revised several times, with NBM, version 4.1 (NBM v4.1), using an updated layer snow ratio versus temperature polynomial curve and an updated layer weighting factor. The Cobb polynomial curve used in NBM v4.1 determines the average SLR at each layer in a profile by adding and subtracting one to the layer temperature (°C) to account for temperature gradients in near-freezing conditions. The updated layer weighting factor uses the square root

TABLE 1. Descriptions of each SLR prediction method.

Method	Description
RF	0.5, 1, and 2 km AGL temperature and wind fed into RF machine learning algorithm
Cobb	925–300-mb layer snow ratio, vertical velocity, and relative humidity (used in NBM v4.1)
MaxTAloft	Maximum temperature between 2000 ft AGL and 400 hPa (used in NBM v4.1)
50% Blend	50/50 blend of Cobb and MaxTAloft

TABLE 2. The 2 × 2 contingency table used for verification.

		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

of layer vertical velocity (cm s^{-1}). Prior to NBM v4.1, the polynomial curve was developed based on results from Dubé (2003), Baxter et al. (2005), and Ware et al. (2006) (The COMET Program 2023). The Roebber et al. (2003) SLR method uses 6-h QPF, surface wind speed, a surface compaction value dependent on month, and temperature and relative humidity interpolated to 14 sigma levels fed into an ensemble of 10 artificial neural networks. The 850–700-mb thickness method is based on a linear relationship between SLR and 850–700-mb thickness derived from 63 mid-Atlantic cases (The COMET Program 2023). The efficacy of these methods over the western CONUS is largely undocumented.

d. SLR methods applied to GFS and HRRR forecasts

The RF, MaxTAloft, and Cobb algorithms were then applied to generate SLR forecasts and QSFs from GFS and HRRR profile forecasts initialized at 0000 and 1200 UTC. The MaxTAloft and Cobb algorithms were selected because they are used in the operational NBM. Both iterations of the MaxTAloft and Cobb algorithms in this study are used operationally in NBM v4.1. A 50/50 blended forecast based on MaxTAloft and Cobb was also evaluated as it is applied to HRRR forecasts in the NBM. Table 1 summarizes each of these SLR methods. We also evaluated the Roebber method, which is based on Roebber et al. (2003) and was designed to predict probabilities for three snow density classes (light, average, and heavy). However, the NBM converts these probabilities into an explicit SLR, and we found that conversion resulted in anomalously high SLRs and poor performance at CLN. Thus, we have not included the Roebber method or the 33% blends that use it. Due to the absence of submodel terrain in the BUFKIT profiles, the 850–700-hPa thickness method was not evaluated as it requires the use of data below model topography.

The RF, MaxTAloft, and Cobb algorithms were applied to GFS and HRRR QPF and profile forecasts for Alta that were downloaded from the Iowa State University BUFKIT Warehouse (https://meteor.geol.iastate.edu/~ckarsten/bufkit/data/). These BUFKIT forecasts are extracted from the ~13-km GFS and ~3-km HRRR at 40.58°N, 111.63°W, which is ~875 m northeast of CLN, are available at hourly intervals, are substantially smaller files, and provide higher vertical resolution than standard pressure-level model output grids. The high temporal and vertical resolutions of the BUFKIT profiles allow forecasters at the Salt Lake City NWS Forecast Office to receive a more complete assessment of the atmosphere when generating forecasts over complex terrain such as LCC. The GFS and HRRR temperature and wind profiles were interpolated to the levels needed for each algorithm. Ground level

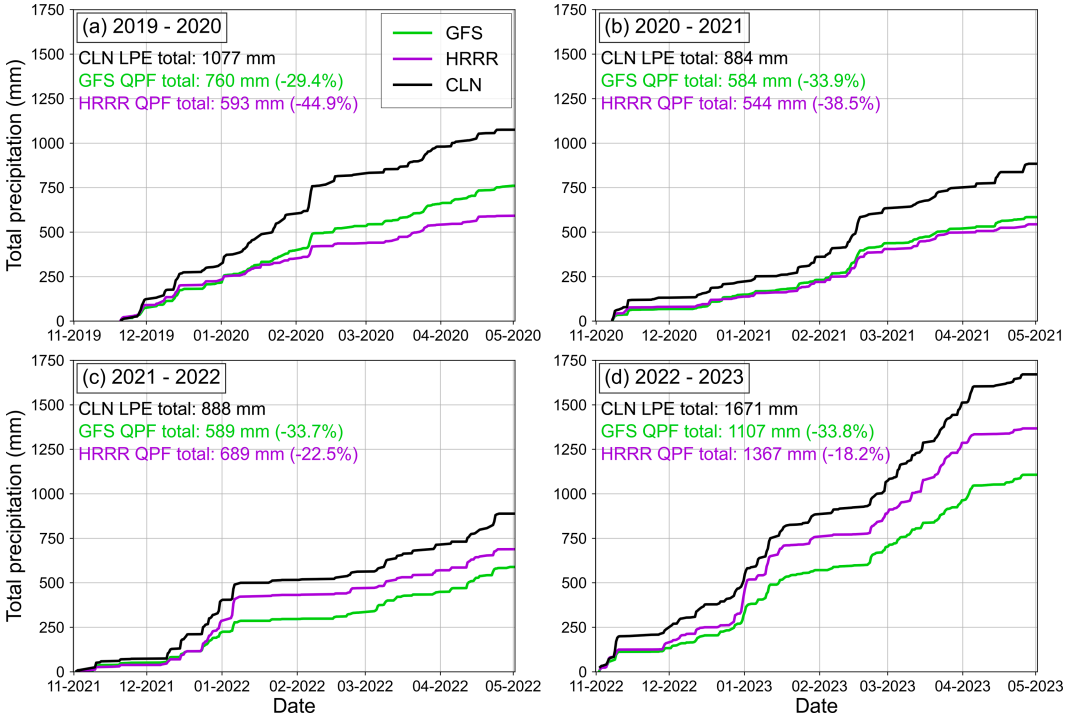

FIG. 3. Observed and forecast accumulated cool-season precipitation and annotated cool-season precipitation totals and seasonal QPF biases as percentages at CLN during the (a) 2019/20, (b) 2020/21, (c) 2021/22, and (d) 2022/23 cool seasons.

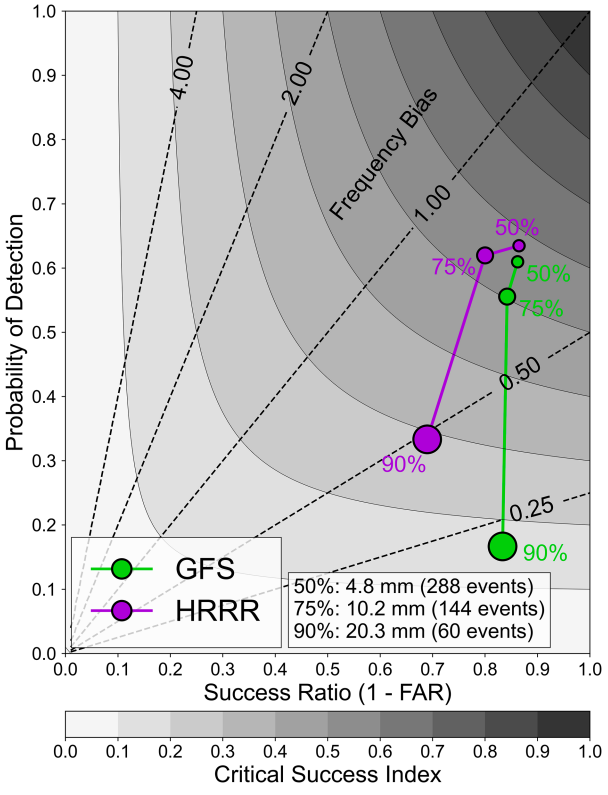

FIG. 4. Performance diagram of GFS and HRRR 12-h QPF skill relative to 50th, 75th, and 90th percentile LPE events (values and number of events annotated at bottom) at CLN for the 2019/20–2022/23 cool seasons. Circle sizes increase with observed LPE percentiles. Shaded contours indicate CSI values, and dashed line contours are frequency bias thresholds.

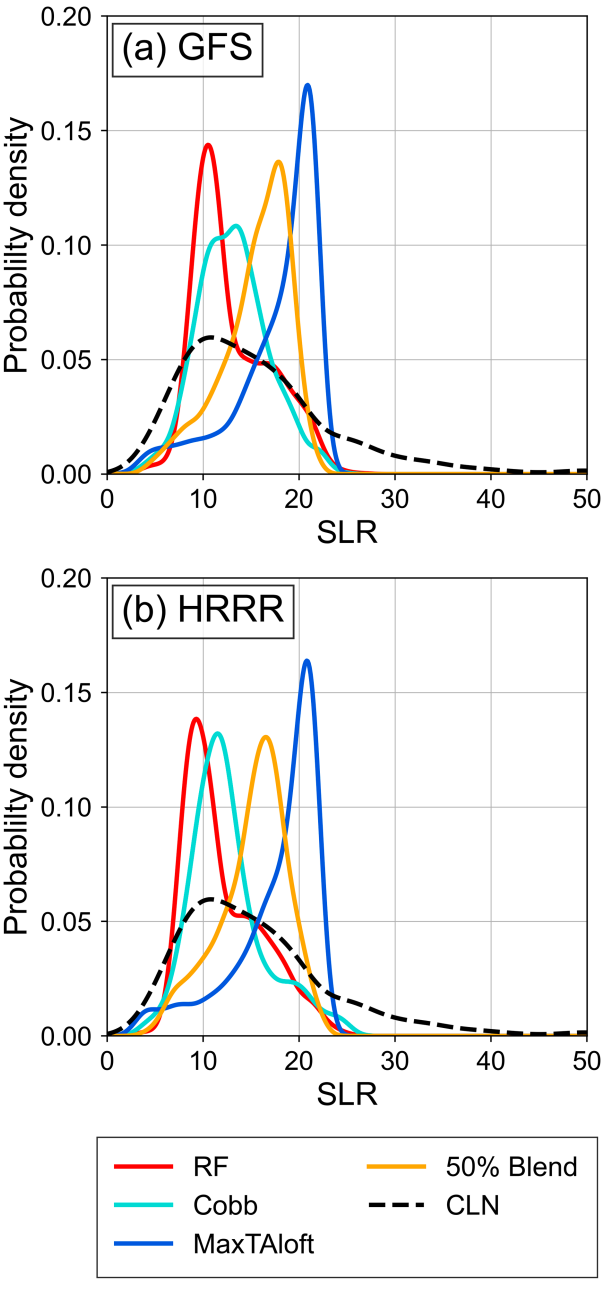

FIG. 5. PDFs of observed and forecast SLR from each SLR method derived from the (a) GFS and (b) HRRR.

(i.e., 0 m AGL) was defined based on the elevation of CLN and not the model topography to utilize a profile that matched as closely as possible to that expected above the station elevation. Any model levels below CLN's elevation were removed. These levels varied with GFS and HRRR updates.

For SLR calculations, the height of the highest forecast 0°C wet-bulb temperature level was assumed to be the top of the melting layer. If CLN was above this level, the forecast SLR was used directly. If CLN was more than 300 m below this level, the precipitation was assumed to be rain. However, there were no soundings in which this was the case during the study period. When CLN was below the 0°C wet-bulb temperature level and in the melting layer, the forecast SLR was adjusted following

$$\mathrm{SLR} = \mathrm{SLR}_{\mathrm{init}} \times \left(\frac{Z_{\mathrm{CLN}} + \mathrm{ML}_{\mathrm{thick}} - Z_{\mathrm{wbz}}}{\mathrm{ML}_{\mathrm{thick}}} \right), \qquad (2)$$

where $\mathrm{SLR}_{\mathrm{init}}$ is the initial predicted SLR, Z_{CLN} is CLN's elevation (2945 m MSL), $\mathrm{ML}_{\mathrm{thick}}$ is the melting layer depth (300 m), and Z_{wbz} is the height (m MSL) of the highest 0°C wet-bulb temperature level. The 300-m melting layer depth is based on White et al. (2010). Essentially, this linearly reduces the forecast SLR to 0 over a depth of 300 m below the 0°C wet-bulb temperature level. This is a simple approach that was applied to the SLR methods and was used during 11 SLR events. Although a new melting SLR technique was implemented into the NBM v4.2 for surface wet-bulb temperatures > 0°C (Leone et al. 2023), we did not incorporate it into our methods as CLN infrequently experiences surface temperatures above freezing during cool-season precipitation events.

QSF for each 12-h CLN observing period was based on hourly QPFs multiplied with hourly SLRs and integrated over the 12-h period. This produced slightly better results than

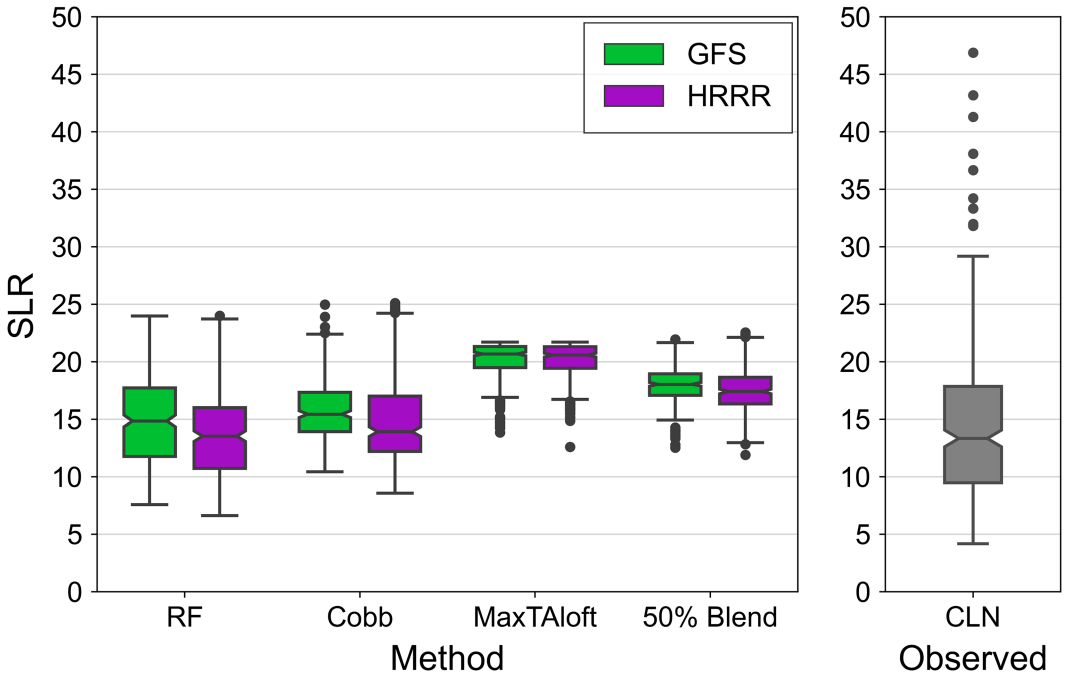

FIG. 6. Box-and-whisker chart of forecast SLR for each method derived from the (left) GFS and HRRR and (right) observed SLR at CLN. The horizontal line in each box represents the median, the box represents the interquartile range, the whiskers represent the 5th and 95th percentiles, and the filled circles represent outliers. Notches indicate the 95% confidence interval of the median.

multiplying the 12-h QPF with a 12-h mean SLR based on 12-h mean predictor variables. Then, the 12-h SLR, SLR_{12h}, was calculated using

$$SLR_{12h} = \frac{QSF_{12h}}{QPF_{12h}}, \tag{3}$$

where QSF_{12h} is the 12-h QSF and QPF_{12h} is the 12-h QPF. If, however, the 12-h QPF was <0.254 mm or the QPF was 0, a time-averaged SLR was used based on the mean of 1-h SLRs during the 12-h period. This enables a comparison of model-derived SLR with observed SLR during periods that there was no QPF, but snowfall was observed. For brevity, the 12h subscripts above are dropped hereafter.

e. *Verification*

Using 0000 and 1200 UTC model initializations for verification, forecast–observation pairs were created by matching forecast period end times with observation collection times. Since observations were collected in local time, 1100–2300 UTC or 1000–2200 UTC forecasts from the 0000 UTC initialized GFS and HRRR were validated during standard or daylight-saving time, respectively. Likewise, 2300–1100 UTC or 2200–1000 UTC forecasts were validated for 1200 UTC initialized forecasts. For both initializations, the two forecast periods correspond to validating forecast hours 11–23 or 10–22 during standard or daylight-saving time, respectively. These forecast hours are used as they are the earliest available that align with each observation period. Out of 1451 possible 12-h forecast–observation pairs, 1447 were evaluated based on the availability of both the GFS and HRRR. We did not evaluate forecasts at other initialization times.

Finally, QPF and QSF were evaluated using a 2 × 2 contingency table (Table 2 following Mason 2003) with hit rate (HR) defined as

$$HR = \frac{a}{a + c}, \tag{4}$$

false alarm ratio (FAR) defined as

$$FAR = \frac{b}{a + b}, \tag{5}$$

and critical success index (CSI) defined as

$$CSI = \frac{a}{a + b + c}, \tag{6}$$

where a is the number of hits, b is the number of false alarms, and c is the number of misses. A hit occurs when a forecast and the corresponding observation exceed the specified threshold; a false alarm occurs when the forecast exceeds the specified threshold, but the corresponding observation does not, and a miss occurs when the forecast does not exceed the specified threshold, but the corresponding observation does. HR (also known as the probability of detection) is the fraction of observed events correctly forecasted, FAR is the fraction of forecasted events that did not occur, and CSI (also known as threat score) measures the fraction of observed and

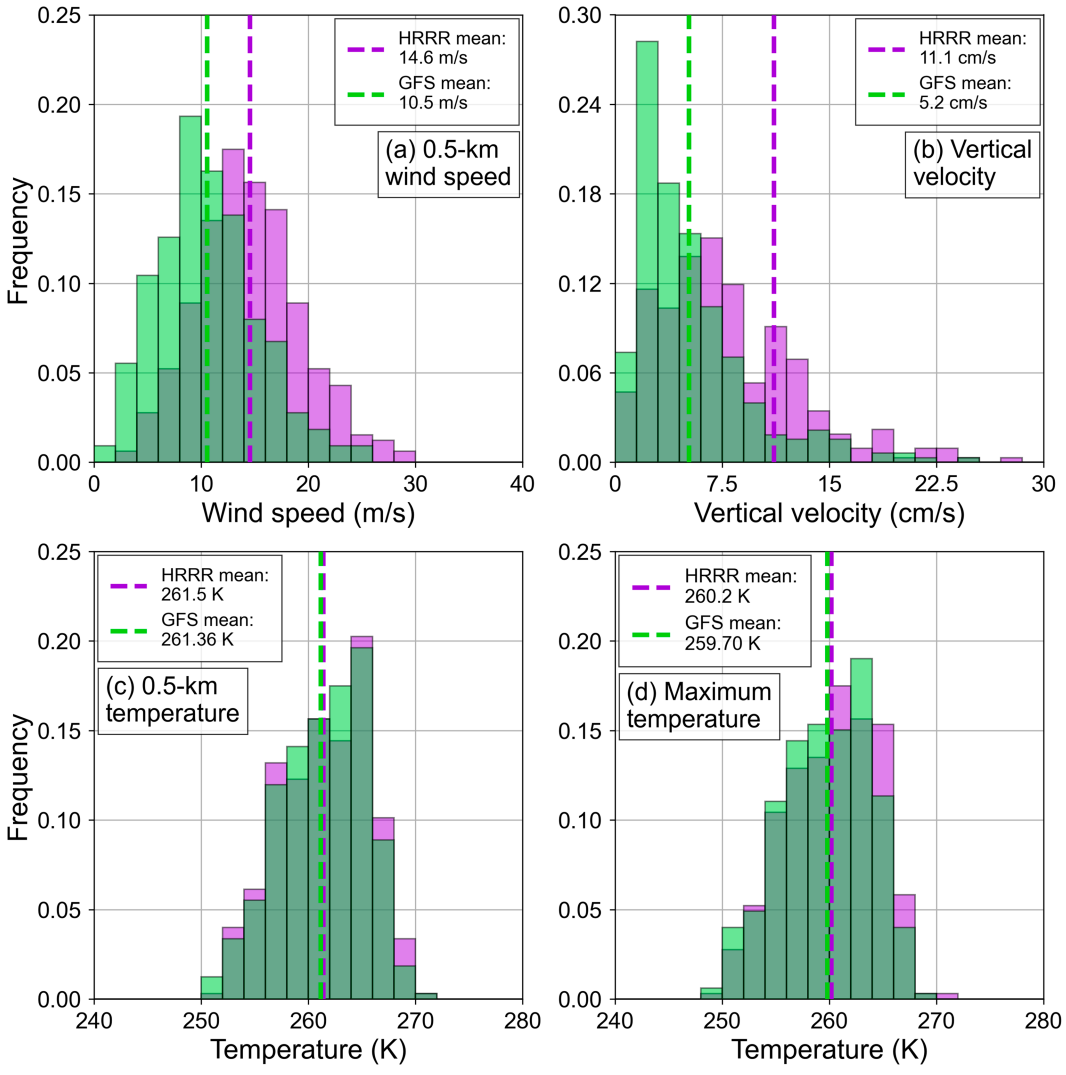

FIG. 7. Dual-colored 1D histograms of GFS (green) and HRRR (light purple) (a) 0.5-km wind speed (m s^{-1}), (b) vertical velocity (cm s^{-1}) calculated as in NBM v4.1 for the Cobb method, (c) 0.5-km temperature (K), and (d) maximum temperature between 2000 ft (610 m) AGL and 400 hPa. Each variable is averaged during each 12-h SLR event. Vertical lines are population means for each model with values annotated.

forecasted events that are correctly predicted but adjusted for hits related to random chance (Wilks 2011). These verification metrics are presented in performance diagrams, which illustrate the accuracy and bias in one diagram presenting HR, success ratio (1 − FAR), and CSI (Roebber 2009). We use the 50th, 75th, and 90th percentile observed LPE and snowfall events shown in Figs. 2d–f as the thresholds in the performance diagrams.

3. QPF validation

a. Seasonal accumulations

Time series of accumulated observed and forecast precipitation illustrate the QPF biases that occur each cool season of the validation period (Figs. 3a–d). Despite differing horizontal grid spacings and model architectures, the GFS and HRRR underpredict seasonal LPE at CLN in all four cool seasons with relatively similar overall mean magnitudes of 33% and 29%, respectively. Internal testing revealed a similar underprediction of cool-season precipitation at nearby GFS and HRRR native grid points (not shown). These results are consistent with Gowan et al. (2018) who found the GFS and HRRR generally underpredicted cool-season precipitation at mountain sites over the western CONUS during the 2016/17 cool season. The largest GFS LPE underprediction (33.9%) occurred during the 2022/23 cool season (Fig. 3d), whereas the largest HRRR LPE underprediction (44.9%) occurred during the 2021/22 cool season (Fig. 3c). Prior to the 2021/22 cool season, the GFS was wetter than the HRRR (Figs. 3a,b), but beginning with the 2021/22 cool season, the HRRR was wetter than the GFS (Figs. 3c,d). Upgrades to the HRRR in

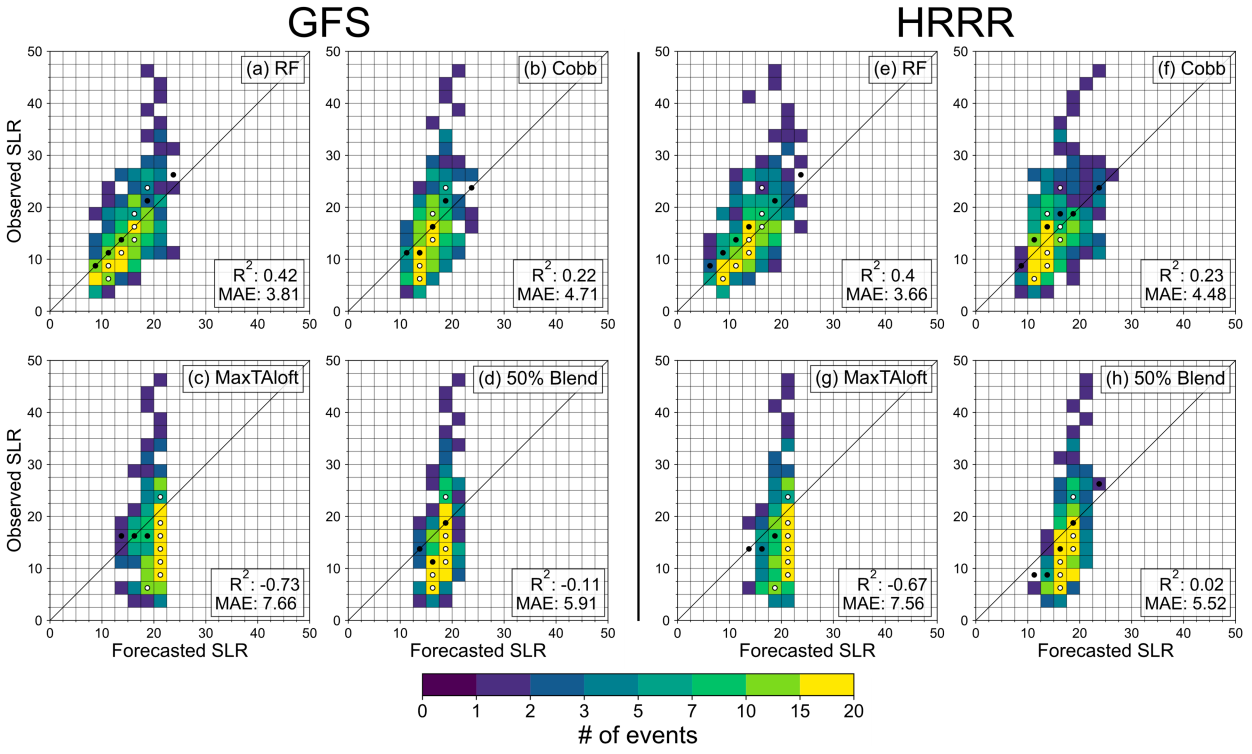

FIG. 8. Bivariate histograms of forecast and observed SLR at CLN for (a) RF, (b) Cobb, (c) MaxTAloft, and (d) 50% Blend derived from the GFS during 2019/20–2022/23 cool seasons with coefficient of determination R^2 and MAE annotated in the lower right. (e)–(h) As in (a)–(d), but derived from the HRRR. White dots are the median observed event size in each bin, and black dots are the forecast event size in each bin. Dots are not shown for bins with fewer than 10 events. The bin width is set to 2.5.

December 2020 (NCEP 2020) and the GFS in March 2021 (NCEP 2021) may have affected model biases, especially for the HRRR which appeared to result in a smaller dry bias. The majority of the seasonal QPF bias in both models reflected the underforecasting of a few large LPE events. For example, a multiday storm from 5 to 7 February 2020 produced 139.7 mm of LPE, whereas the GFS and HRRR produced 79.9 and 58.7 mm, underforecasting precipitation by 42.8% and 58%, respectively.

b. *Individual events*

Using statistical measures based on a standard 2×2 contingency table (Table 2), we examined QPF performance during individual events relative to observed 50th (4.8 mm), 75th (10.2 mm), and 90th (20.3 mm) LPE percentiles during the 2019/20–2022/23 validation period using performance diagrams (Roebber 2009). Points above the 1.0 frequency bias line indicate an overprediction of LPE event frequency, whereas points below the 1.0 frequency bias line indicate an underprediction of LPE event frequency. Forecast accuracy increases toward the top right of the diagram where the probability of detection (POD), success ratio (SR), and CSI are all 1. The GFS and HRRR both exhibit frequency biases < 1 for all three LPE percentiles (Fig. 4), indicating overall negative QPF biases consistent with the seasonal LPE underprediction (e.g., Figs. 3a–d). Underforecasting is greatest for the 90th percentile events. Despite differing model architectures and horizontal grid spacings, the GFS and HRRR produce comparable SRs and CSIs for 50th and 75th percentile LPE events but slightly different PODs for 75th percentile LPE events (0.55 and 0.61, respectively). For 90th percentile LPE events, the GFS and HRRR exhibit a sharp decrease in accuracy, particularly in CSI and SR, with the HRRR performing slightly better than the GFS. These results illustrate that for QPF at Alta, the GFS and HRRR exhibit relatively similar accuracy for modest 50th and 75th percentile LPE events and degraded performance for the largest and often highly impactful 90th percentile events, with the higher-resolution HRRR exhibiting somewhat smaller dry biases and greater accuracy for the 90th percentile events.

4. SLR validation

a. *Forecast distributions*

We now compare probability density functions (PDFs) of forecasted SLRs from each of the four methods (RF, Cobb, MaxTAloft, and 50% Blend) with observed SLR during the 2019/20–2022/23 evaluation period (Figs. 5a,b). The mode of the RF SLR derived from the GFS best matches observed, but the distribution is narrower due to an underprediction of low ($\lesssim$8) and high ($\gtrsim$22) SLRs (Fig. 5a). The Cobb SLR distribution is broader than the RF, but its mode is shifted toward higher SLRs and the underprediction of low and high

SLRs is still evident. The MaxTAloft distribution is highly skewed to higher SLRs with an unrealistic mode of 21, which, when combined with Cobb, leads to an unrealistically high SLR mode of 18 for the 50% Blend.

Compared to the GFS-derived SLRs, the HRRR-derived RF and Cobb SLR modes are shifted toward slightly lower values but remain close to the observed SLR mode (Fig. 5b). The distributions remain too narrow with the HRRR-derived Cobb distribution noticeably narrower than its GFS-derived distribution. MaxTAloft and 50% Blend have distributions similar to their GFS-derived distributions, with their modes at nearly identical, unrealistically high values. Like the GFS-derived SLRs, all the HRRR-derived methods rarely produce SLRs $\gtrsim 25$.

Contrasts between the GFS- and HRRR-derived RF and Cobb SLRs are more clearly illustrated by box-and-whisker plots (Fig. 6). For both RF and Cobb, the HRRR-derived distributions are clearly shifted toward lower SLRs. In contrast, the distributions for MaxTAloft are nearly identical. The HRRR-derived 50% Blend is shifted toward lower values due to the influence of Cobb. The tendency of the RF and Cobb methods to forecast lower HRRR-derived SLRs is due to the differing wind and vertical velocity distributions produced by the GFS and HRRR at Alta during SLR events. For example, the distribution of the HRRR 0.5-km wind speed averaged during SLR events is shifted toward higher values compared to the GFS, which leads to lower RF SLRs (Fig. 7a; other wind levels used by the RF exhibit similar results). The HRRR vertical velocity distribution is also shifted to higher values, which leads to lower Cobb SLRs (Fig. 7b). In contrast, the GFS and HRRR distributions of 0.5-km temperature and maximum temperature between 2000 ft AGL and 400 hPa are relatively similar, so there are small differences in the MaxTAloft SLR forecasts (Figs. 7c,d).

In summary, the RF and Cobb methods produce GFS- and HRRR-derived forecast SLR distributions with modes closest to observed. MaxTAloft tends to forecast unrealistically high SLRs, which results in the 50% Blend SLRs being too high. All methods produce distributions that are too narrow, fail to predict the highest SLR events ($\gtrsim 25$), and underpredict low SLR events ($\lesssim 8$). Differences in the GFS- and HRRR-derived RF and Cobb SLRs reflect contrasts in the wind and vertical velocity distributions during SLR events produced by the two models, whereas similarities in GFS- and HRRR-derived MaxTAloft and 50% Blend SLRs are a result of nearly identical 0.5-km and maximum temperature distributions during SLR events.

b. *Individual events*

Bivariate histograms further illustrate the biases and accuracy of the SLR forecasts (Fig. 8). Underprediction or overprediction is indicated by frequent event pairs falling above or below the 1:1 line in Fig. 8, respectively, while limited scatter signals accuracy. For the GFS- and HRRR-derived SLR forecasts, the RF has relatively high accuracy for event pairs $\lesssim 20$, the highest overall R^2 values (0.42 and 0.4), and the lowest MAEs (3.81 and 3.66) (Figs. 8a,e). Accuracy degrades for

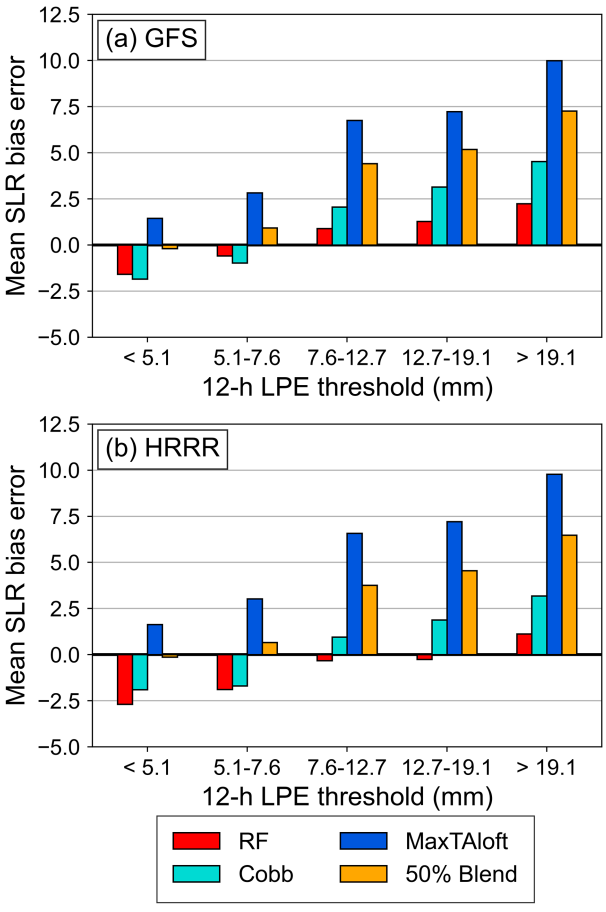

FIG. 9. Mean SLR bias error binned by 12-h LPE thresholds (mm) for the (a) GFS and (b) HRRR.

SLR events $\gtrsim 20$. Cobb performs slightly worse with R^2 values of 0.22 and 0.23 and MAEs of 4.71 and 4.5 (Figs. 8b,f). The tendency of MaxTAloft to produce anomalously high SLRs near the aforementioned mode ~21 is evident in the large number of event pairs with a forecast SLR of 20–22.5 (Figs. 8c,g) and is a result of the SLR versus temperature polynomial curve used in the NBM v4.1 MaxTAloft calculations, which maximizes at 22 (The COMET Program 2023). The polynomial curve used by MaxTAloft was developed for environments that feature warm air aloft which rarely occurs at CLN (Fig. 7d), resulting in infrequent SLRs < 15. This leads to negative R^2 values (-0.73 and -0.68) and high MAEs (7.65 and 7.57), indicating that MaxTAloft exhibits worse performance than the SLR climatological mean at Alta (13.5; performance not shown). The performance of MaxTAloft is so poor that even after averaging with Cobb, 50% Blend still yields negative and near-zero R^2 values (-0.11 and 0.02) and exhibits relatively high MAEs (5.9 and 5.53; Figs. 8d,h).

The analysis above is independent of LPE amount. SLR errors will yield larger absolute snowfall errors with increasing LPE. To evaluate how SLR performance varies with LPE, we calculated mean SLR biases for five observed LPE bins (<5.1, 5.1–7.6, 7.6–12.7, 12.7–19.1, and >19.1 mm). For the

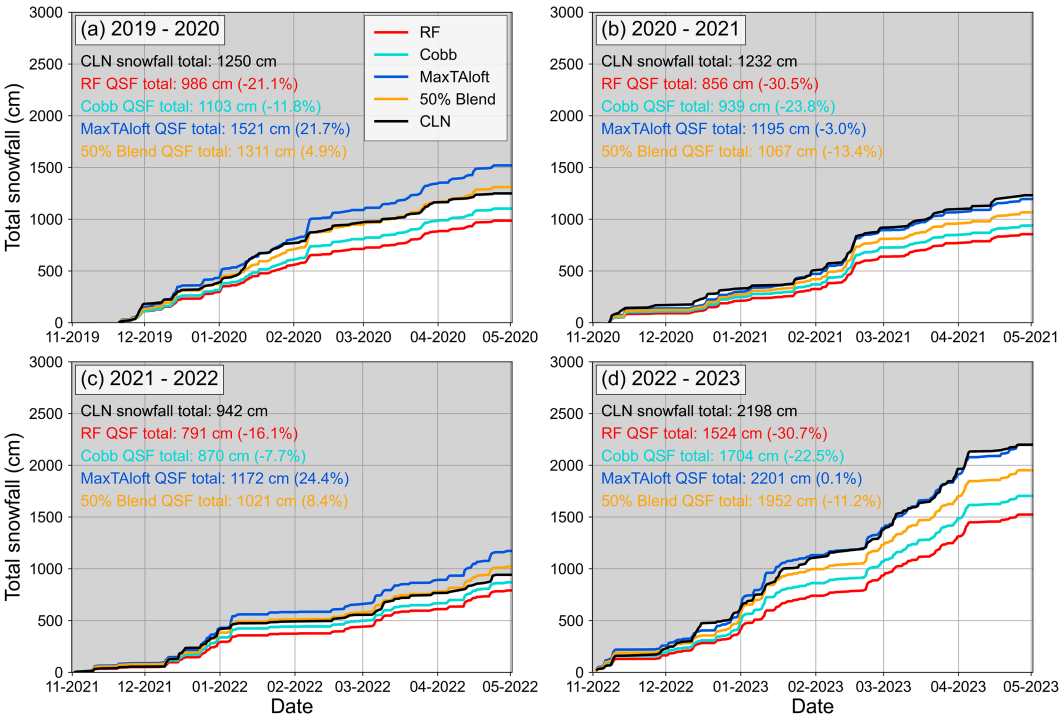

FIG. 10. Observed seasonal snowfall and GFS seasonal QSFs derived from each SLR method at CLN during the (a) 2019/20, (b) 2020/21, (c) 2021/22, and (d) 2022/23 cool seasons. Light gray shading indicates values above CLN snowfall, and white shading indicates values below CLN snowfall. QSF seasonal totals and QSF biases are annotated as percentages in the upper left.

GFS- and HRRR-derived SLR forecasts, RF and Cobb exhibit increasing bias errors with LPE from relatively small negative values ($|SLR| \lesssim 2.5$) for low (<7.6 mm) LPE events to positive values for large LPE events (Figs. 9a,b). This indicates that RF and Cobb tend to forecast slightly lower SLR than observed for small LPE events but higher SLR than observed for large LPE events. For large LPE events, the GFS- and HRRR-derived RF biases are lowest and, in the case of the HRRR, nearly 0 for 7.6–19.1-mm LPE events and $\lesssim 2.5$ for >19.1-mm LPE events, showcasing its relatively low bias during large LPE events. Conversely, GFS- and HRRR-derived MaxTAloft and 50% Blend forecasts exhibit positive mean SLR biases that increase with LPE and are considerably larger than those exhibited by RF and Cobb. The positive bias trend with LPE may reflect the lack of LPE as a predictor for all the methods given the tendency for LPE to negatively correlate with SLR, especially in larger storms (Judson and Doesken 2000; Roebber et al. 2003; Ware et al. 2006; Alcott and Steenburgh 2010).

Collectively, these results illustrate that the RF produces the most accurate SLR forecasts, followed by Cobb, although forecasts of high SLR events are poor for both methods. While Cobb forecasts relatively accurate SLRs compared to the other NBM methods, its SLR forecasts exhibit substantially lower R^2 values and higher MAEs than the RF, revealing its poor performance compared to the RF. The poor performance of MaxTAloft suggests it should not be used for snowfall forecasting at this location. Even after blending MaxTAloft with Cobb's relatively accurate SLR forecasts, the 50% Blend's performance is also poor due to MaxTAloft's anomalously high, unrealistic SLRs. The better performance of the RF and Cobb SLR methods is further evidenced by their relatively small mean SLR biases for large (>12.7 mm) LPE events, while MaxTAloft and 50% Blend produce large mean SLR biases for large LPE events. Thus, the RF, followed by Cobb, produces the most accurate SLR forecasts during high-impact LPE events at Alta.

5. QSF validation

a. *Seasonal accumulations*

Similar to Figs. 3a–d, we present a time series of accumulated observed and forecast snowfall to illustrate the QSF biases that occur each cool season of the validation period (Fig. 10). Despite the RF and Cobb SLR methods producing relatively low SLR biases and MAEs during individual events, they consistently produce negative seasonal QSF biases when applied to the dry-biased GFS QPFs over each cool season. Seasonal QSF derived with the RF underpredicts seasonal snowfall the most during the 2022/23 cool season (30.7%; Fig. 10d), while seasonal QSF derived with Cobb underpredicts seasonal snowfall the most (23.8%) during the 2020/21 cool season (Fig. 10b). Conversely, seasonal QSFs derived from MaxTAloft and 50% Blend exhibit smaller underprediction or even slightly positive seasonal QSF biases. For example, the MaxTAloft-derived seasonal

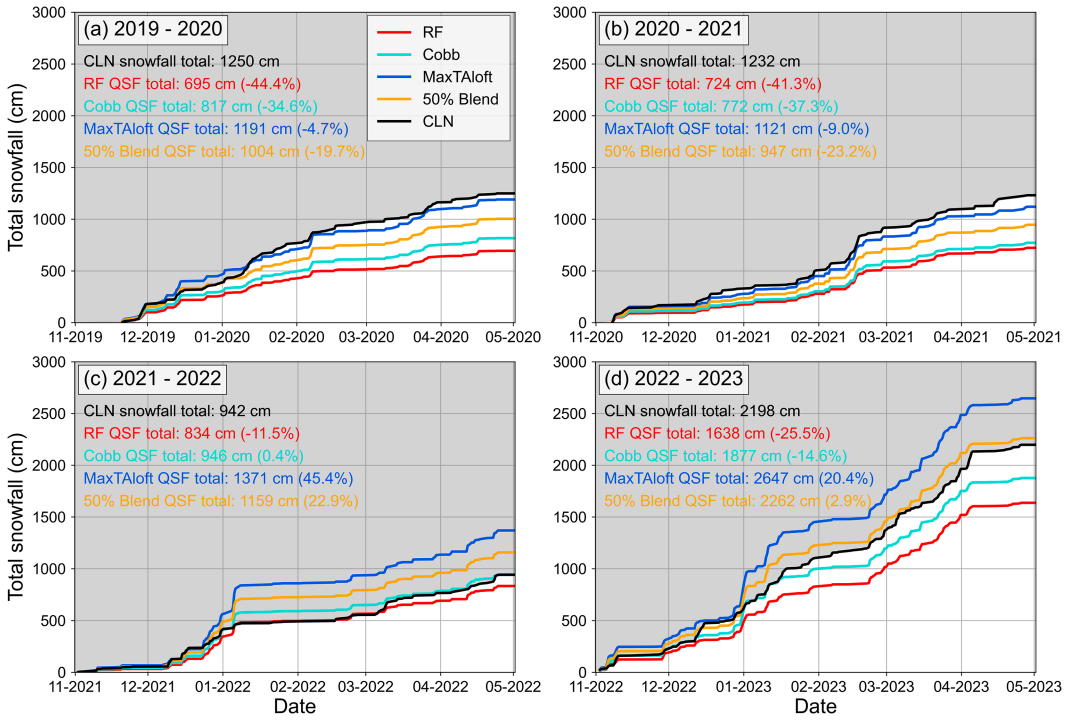

FIG. 11. As in Fig. 10, but for HRRR QSFs.

QSF is very similar to observed during the 2022/23 cool season (Fig. 10d), and the seasonal QSF derived from 50% Blend overpredicts seasonal snowfall by 5.5% during the 2019/20 cool season (Fig. 10a). These results illustrate that the large positive SLR biases exhibited by the MaxTAloft and 50% Blend methods during individual events (Figs. 8c,d,g,h) offset the GFS negative QPF bias (Figs. 3a–d and 4), leading to more accurate seasonal QSFs than those produced by the RF and Cobb methods despite their smaller SLR biases.

For the HRRR, seasonal QSFs derived with RF and Cobb underpredict seasonal snowfall each cool season except during the 2021/22 cool season when the Cobb-derived seasonal QSF overpredicts seasonal snowfall by 0.4% (Fig. 11). Seasonal QSFs derived from MaxTAloft produce small biases during the 2019/20 and 2020/21 cool seasons but higher than observed during the 2021/22 and 2022/23 cool seasons (Figs. 11c,d) with 45.6% and 20.4% overpredictions during each cool season, respectively. These trends reflect the decline in the HRRR dry bias during the latter two cool seasons (Figs. 3c,d). The 2021/22 cool season is the only cool season when RF- and Cobb-derived QSFs both exhibit smaller seasonal QSF biases than the MaxTAloft- and 50% Blend–derived QSFs (Fig. 11c).

b. *Individual events*

We now assess the performance of QSFs for individual events derived from each SLR method for the GFS (Fig. 12). For 50th (7.6 cm) percentile snowfall events, the QSFs derived from each method exhibit frequency biases < 1 and produce comparable CSIs ($\sim$0.55), indicating modest performance and overall underprediction of median snowfall events. Consistent with the above seasonal QSF performance analysis, underprediction, as reflected by a smaller frequency bias, is greatest for RF and Cobb and smallest for MaxTAloft and 50% Blend. PODs and CSIs decline for all four methods as event size increases to 75th percentile (15.2 cm) and 90th percentile (25.4 cm) snowfall events. These performance declines are, however, smallest for MaxTAloft and 50% Blend which have frequency biases $\sim$ 1. Conversely, RF- and Cobb-derived QSFs experience larger decreases in POD and CSI. The better performances of the MaxTAloft- and 50% Blend–derived QSFs for 75th and 90th percentile snowfall events reflect their large positive SLR biases, which offset the GFS dry bias.

Although HRRR QSFs exhibit similar performance as the GFS QSFs for 50th percentile snowfall events, they experience a greater decline in all performance metrics (CSI, SR, and POD) for 75th and 90th percentile snowfall events (Fig. 13). Except for the MaxTAloft-derived QSFs, there is a clear tendency for the HRRR QSFs to underpredict event frequencies, similar to the GFS QSFs. MaxTAloft-derived QSFs show a slight overprediction (frequency biases > 1) of 75th and 90th percentile snowfall events and larger CSIs than the QSFs derived from the other SLR methods. Cobb-derived QSFs perform slightly better than the RF-derived QSFs though both perform poorly for events greater than the 75th percentile. Like the GFS-derived QSFs, HRRR MaxTAloft- and 50% Blend–derived QSFs perform better than the RF- and Cobb-derived QSFs (higher CSI, frequency biases closer to 1) due to their large positive SLR biases.

We calculate QSF mean bias errors (MBEs) derived from each SLR method relative to LPE amounts (same thresholds

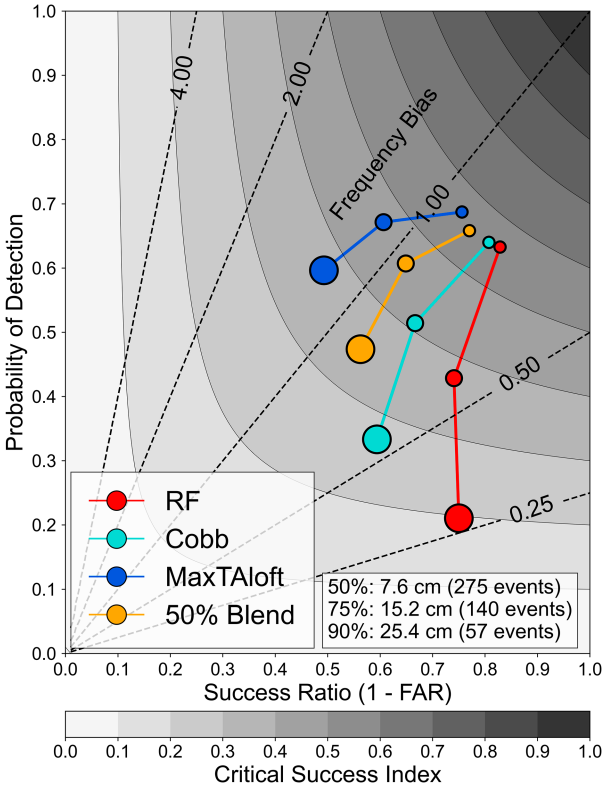

FIG. 12. Performance diagram of GFS 12-h QSF skill relative to 50th, 75th, and 90th percentile observed snowfall events (totals and number of events annotated at bottom). Circle sizes increase with observed snowfall percentiles. Shaded contours indicate CSI values, and dashed line contours are frequency bias thresholds.

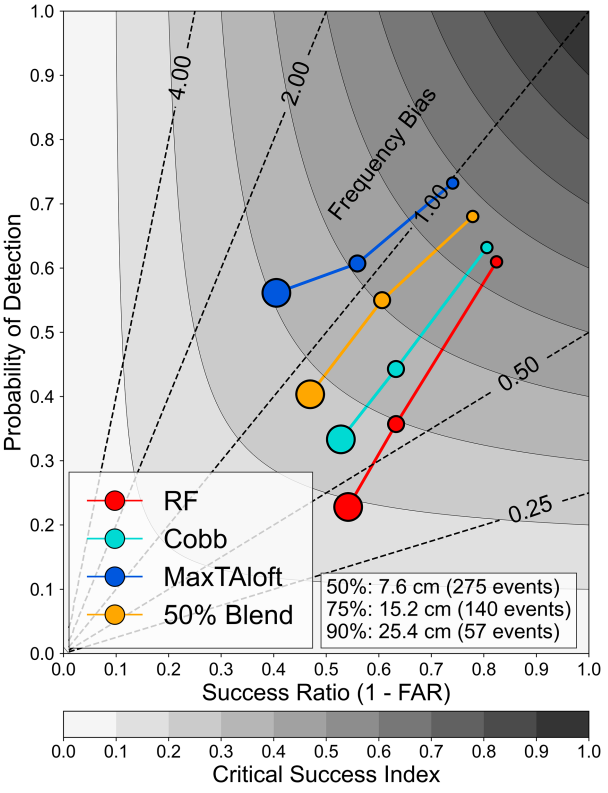

FIG. 13. As in Fig. 12, but for HRRR QSFs.

as in Figs. 9a,b) for the GFS and HRRR to determine when the largest QSF bias errors occur for different LPE event sizes (Figs. 14a,b). For the GFS and HRRR QSFs, the MBEs are positive for LPE events < 5.1 mm but become negative with increasing LPE, indicating a negative decrease in QSF bias error as LPE increases. However, the HRRR QSFs all exhibit considerably larger positive MBEs (>100 cm) than the GFS for LPE events < 5.1 mm, indicating much larger snowfall forecasts than observed. MaxTAloft- and 50% Blend–derived QSFs exhibit the largest MBE for LPE events < 7.6 mm but exhibit the smallest MBE for LPE events ≥ 12.7 mm. Conversely, RF- and Cobb-derived QSFs exhibit the smallest MBE for LPE events < 7.6 mm and the largest MBE for LPE events ≥ 12.7 mm. While the MBEs exhibited by each QSF are relatively similar; overall, the HRRR tends to produce larger QSF MBEs than the GFS.

We find different mean QSF MBEs when MBEs are instead calculated for five observed SLR bins (<10, 10–15, 15–20, 20–25, and >25). GFS and HRRR QSFs exhibit negative MBEs with increasing SLRs with somewhat larger MBEs exhibited when derived from the HRRR, similar to Figs. 14a and 14b (Figs. 15a,b). In contrast to Figs. 14a and 14b, MaxTAloft QSFs produce the smallest MBE for all SLR event thresholds except for SLR events < 10 for the GFS and HRRR. Meanwhile, RF- and Cobb-derived QSFs produce the largest MBE for all SLR bins except for SLR events < 10. Like Figs. 10–13, these results indicate that the dry biases present in the GFS and HRRR offset the positive SLR biases exhibited by the MaxTAloft SLR method, leading to smaller QSF MBEs than the more accurate RF and Cobb methods during some SLR events. High SLR events often contain a smaller amount of LPE than low SLR events (Alcott and Steenburgh 2010, see their Fig. 5f) which can alter MBEs exhibited by the QSFs. Thus, due to MaxTAloft- and 50% Blend–derived QSFs' reliance on QPF underprediction for accuracy, they exhibit smaller QSF MBEs than RF- and Cobb-derived QSFs during SLR events ≥ 15 for the GFS and HRRR.

Despite the large positive SLR biases exhibited by MaxTAloft and 50% Blend, we find that they tend to produce smaller seasonal QSF biases than the better-performing RF and Cobb SLR methods by offsetting the dry biases present in the GFS and HRRR. The overall better performances of MaxTAloft- and 50% Blend–derived QSFs are evident in the performance diagrams by their frequency biases closer to 1 and higher CSIs for all event sizes, which is due to their large positive SLR biases. GFS and HRRR QSF overall performances are similar for 50th percentile snowfall events but diverge for events greater than the 75th percentile although QSF performance declines are greater for the HRRR than the GFS. Bias errors for the RF- and Cobb-derived GFS and HRRR QSFs are lesser during small LPE and SLR events but are greater during large LPE and SLR events. Conversely, the large positive SLR biases

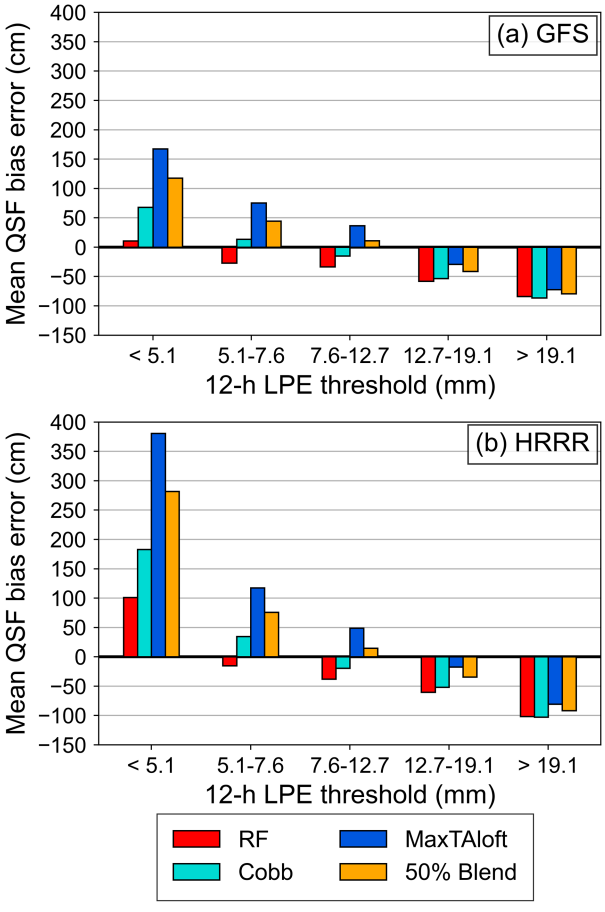

FIG. 14. Mean QSF bias error (cm) binned by 12-h LPE thresholds (mm) for the (a) GFS and (b) HRRR.

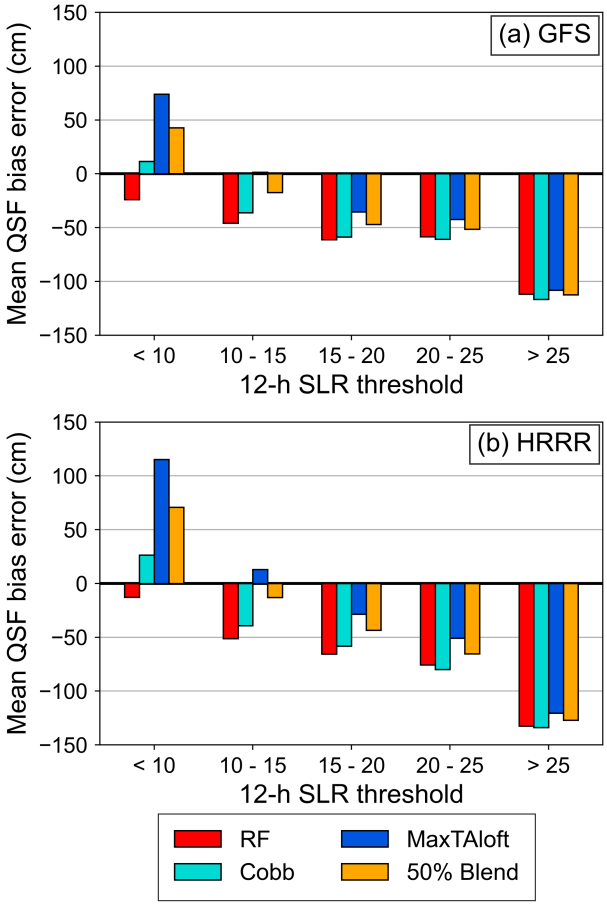

FIG. 15. As in Fig. 14, but binned by 12-h observed SLR thresholds.

exhibited by MaxTAloft and 50% Blend lead to greater QSF bias errors during small LPE and SLR events but lesser bias errors during small LPE and SLR events. Overall, these results indicate a paradox in snowfall forecasting at this location: An SLR forecast with a large positive bias (MaxTAloft and 50% Blend) applied to a QPF with a dry bias produces a more accurate QSF than an SLR forecast with minimal bias (RF and Cobb). Thus, identifying sources of QSF bias error is necessary to improve QSF performance at Alta.

6. Conclusions

This study has examined the performance of QPFs, SLR forecasts, and QSFs produced by or derived from the GFS and HRRR at a high-elevation observing site in upper LCC during the 2019/20–2022/23 cool seasons. Located in the Wasatch Range of northern Utah, LCC frequently experiences high-impact winter storms and avalanche closures and serves as a testbed for evaluating snowfall forecasts due to its high-quality snowfall observations. Both the GFS and HRRR exhibited mean negative QPF biases during the study period with an overall LPE underprediction of 33% and 29%. Model upgrades prior to the 2021/22 cool season appear to have reduced but not eliminated the HRRR dry bias beginning with the 2021/22 cool season.

Despite differing model architectures and horizontal grid spacings, the GFS and HRRR exhibit similar QPF performance (e.g., similar CSI) for 50th and 75th percentile LPE events. For 90th percentile LPE events, both models exhibit a significant decline in accuracy with the HRRR exhibiting the best overall performance. This is consistent with Gowan et al. (2018) who found the HRRR produces more accurate QPFs of large precipitation events than the GFS over interior CONUS mountain ranges.

A newly developed SLR prediction algorithm that uses an RF trained on local data outperforms existing operational SLR forecast methods when derived from either the GFS or HRRR by producing the most accurate SLR forecast distributions, the lowest MAE (3.7), and the highest R^2 value (0.42). Cobb produces less accurate SLR forecasts, and MaxTAloft and its 50% Blend produce the least accurate SLR forecasts as indicated by their unrealistically narrow SLR forecast distributions, MAEs > 7, and negative or near-zero R^2 scores. None of the SLR methods can reliably predict SLR events ≥ 25 when derived from the GFS or HRRR, highlighting the shortcomings of all the methods when forecasting the highest SLR events. Differences in GFS- and HRRR-derived RF and Cobb

SLRs are a result of their differing low-level wind speed and vertical velocity distributions during SLR events. SLR forecasts produced by the RF, followed by Cobb, exhibit the smallest MBE during the often high-impact, large LPE events. Given the large observed variability in SLR at Alta and the different influences on SLR (Pomeroy and Brun 2001; Roebber et al. 2003; Baxter et al. 2005; Cobb and Waldstreicher 2005; Byun et al. 2008; Alcott and Steenburgh 2010), it is unsurprising that the combination of the RF's multiple predictors and its ability to account for nonlinear relationships between SLR and its predictors result in its superior SLR forecast performance at Alta.

Despite more accurate SLR forecasts, QSFs derived with the RF method did not produce the best snowfall forecasts. Instead, the large positive SLR biases exhibited by the MaxTAloft and 50% Blend methods offset the GFS and HRRR dry biases, resulting in more accurate QSFs. The Cobb QSFs also slightly outperform the RF-derived QSFs due to the somewhat larger SLR biases exhibited by Cobb compared to the RF. The performances of the GFS and HRRR QSFs for 50th percentile snowfall events are similar but experience large performance decreases for snowfall events greater than the 75th percentile, with greater performance decreases for these events evident in the HRRR. While MaxTAloft- and 50% Blend–derived GFS and HRRR QSFs produce the largest MBE for small LPE and SLR events, they produce smaller MBE than the RF- and Cobb-derived QSFs for large LPE and SLR events. Collectively through these findings, we reveal a paradox in snowfall forecasting at this location: the most accurate snowfall forecasts are produced from an SLR forecast with a large positive bias that is applied to a QPF with a dry bias. However, as numerical models improve and machine learning approaches are used to reduce QPF bias, the performance of MaxTAloft- and 50% Blend–derived QSFs will decline because their large positive SLR bias will yield QSF overforecasting, rendering this an unsustainable approach. Thus, as previous work by Byun et al. (2008) and Alcott and Steenburgh (2010) emphasized, our results highlight the need to identify sources of QSF bias error to improve QSF performance at Alta.

These findings highlight the potential for machine learning applied to high-quality snowfall observations from snow-safety teams for improved SLR forecasts across the western CONUS as well as the importance of identifying sources of snowfall forecast bias error. The RF consistently outperforms the operational NBM SLR forecast methods but produces the least accurate snowfall forecasts due to the evident dry biases in the GFS and HRRR. Despite the RF's underprediction of snowfall, its superior SLR forecast performance is advantageous during events that are sensitive to SLRs, making it more useful for avalanche mitigation applications (Mueller 2001; Schweizer et al. 2003). A pathway for future work is adding a bias-corrected QPF derived from the GFS and HRRR, which could alter the performance of the QSFs derived from the SLR methods. Furthermore, adding QPF as a predictor to the RF could be beneficial during high LPE events given the tendency for LPE to negatively correlate with SLR (Judson and Doesken 2000; Roebber et al. 2003; Ware et al. 2006; Alcott and Steenburgh 2010). Overall, a more comprehensive examination of QPFs, SLR forecasts, and QSFs at other western CONUS mountain sites could highlight differing performances among the SLR methods and reveal the contributing factors to degraded QSF performance in different snow climates.

Acknowledgments. We thank the University of Utah Center for High Performance Computing for computer support, Alta Ski Area and ski patrol for providing and collecting the observations used in this study, and three anonymous reviewers as well as Michael Wasserstein who helped improve this manuscript. This article is based on research supported by NOAA Weather Program Office Grants NA19OAR4590137 and NA22OAR4590521 and NOAA National Weather Service CSTAR Program Grant NA20NWS4680046. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect those of the National Oceanic and Atmospheric Administration.

Data availability statement. The CLN observations used in this study are openly available and can be found online at the University of Utah Research Data Repository (https://hive.utah.edu/concern/datasets/0r967383v). The BUFKIT data are available from The BUFKIT warehouse 2023. (https://meteor.geol.iastate.edu/~ckarsten/bufkit/bufkit.html). Terrain data used in Fig. 1 are available from ESRI (https://services.arcgisonline.com/arcgis/rest/services/Elevation). The ERA5 reanalysis data are available from the Copernicus Climate Change Service (Hersbach et al. 2018a,b). The random forest SLR algorithm is openly available from GitHub (https://github.com/mdpletcher/SLR_random_forest_pletcher/tree/main).

REFERENCES

Alcott, T. I., and W. J. Steenburgh, 2010: Snow-to-liquid ratio variability and prediction at a high-elevation site in Utah's Wasatch Mountains. *Wea. Forecasting*, **25**, 323–337, https://doi.org/10.1175/2009WAF2222311.1.

——, ——, and N. F. Laird, 2012: Great Salt Lake–effect precipitation: Observed frequency, characteristics, and associated environmental factors. *Wea. Forecasting*, **27**, 954–971, https://doi.org/10.1175/WAF-D-12-00016.1.

Andretta, T. A., and D. S. Hazen, 1998: Doppler radar analysis of a Snake River plain convergence event. *Wea. Forecasting*, **13**, 482–491, https://doi.org/10.1175/1520-0434(1998)013<0482:DRAOAS>2.0.CO;2.

Baxter, M. A., C. E. Graves, and J. T. Moore, 2005: A climatology of snow-to-liquid ratio for the contiguous United States. *Wea. Forecasting*, **20**, 729–744, https://doi.org/10.1175/WAF856.1.

Benjamin, S. G., J. M. Brown, and T. G. Smirnova, 2016: Explicit precipitation-type diagnosis from a model using a mixed-phase bulk cloud–precipitation microphysics parameterization. *Wea. Forecasting*, **31**, 609–619, https://doi.org/10.1175/WAF-D-15-0136.1.

Birk, K., E. Lenning, K. Donofrio, and M. T. Friedlein, 2021: A revised Bourgouin precipitation-type algorithm. *Wea. Forecasting*, **36**, 425–438, https://doi.org/10.1175/WAF-D-20-0118.1.

Blattenberger, G., and R. Fowles, 1995: Road closure to mitigate avalanche danger: A case study for little cottonwood canyon. *Int. J. Forecasting*, **11**, 159–174, https://doi.org/10.1016/0169-2070(94)02008-D.

Bourgouin, P., 2000: A method to determine precipitation types. *Wea. Forecasting*, **15**, 583–592, https://doi.org/10.1175/1520-0434(2000)015<0583:AMTDPT>2.0.CO;2.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Byun, K.-Y., J. Yang, and T.-Y. Lee, 2008: A snow-ratio equation and its application to numerical snowfall prediction. *Wea. Forecasting*, **23**, 644–658, https://doi.org/10.1175/2007WAF2006080.1.

Cobb, D. K., Jr., and J. Waldstreicher, 2005: A simple physically based snowfall algorithm. *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction*, Washington, DC, Amer. Meteor. Soc., 2A.2, http://ams.confex.com/ams/pdfpapers/94815.pdf.

Craven, J. P., D. E. Rudack, and P. E. Shafer, 2020: National blend of models: A statistically post-processed multi-model ensemble. *J. Oper. Meteor.*, **8**, 1–14, https://doi.org/10.15191/nwajom.2020.0801.

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.

Dubé, I., 2003: From mm to cm … study of snow/liquid water ratios in Quebec. Meteorological Services of Canada—Quebec region, Internal Rep., 127 pp., https://meted.ucar.edu/norlat/snowdensity/from_mm_to_cm.pdf.

Espeholt, L., and Coauthors, 2022: Deep learning for twelve hour precipitation forecasts. *Nat. Commun.*, **13**, 5145, https://doi.org/10.1038/s41467-022-32483-x.

Ferber, G. K., C. F. Mass, G. M. Lackmann, and M. W. Patnoe, 1993: Snowstorms over the Puget sound lowlands. *Wea. Forecasting*, **8**, 481–504, https://doi.org/10.1175/1520-0434(1993)008<0481:SOTPSL>2.0.CO;2.

Gerber, F., and Coauthors, 2018: Spatial variability in snow precipitation and accumulation in COSMO–WRF simulations and radar estimations over complex terrain. *Cryosphere*, **12**, 3137–3160, https://doi.org/10.5194/tc-12-3137-2018.

——, R. Mott, and M. Lehning, 2019: The importance of near-surface winter precipitation processes in complex alpine terrain. *J. Hydrometeor.*, **20**, 177–196, https://doi.org/10.1175/JHM-D-18-0055.1.

Gillies, R. R., S.-Y. Wang, and M. R. Booth, 2012: Observational and synoptic analyses of the winter precipitation regime change over Utah. *J. Climate*, **25**, 4679–4698, https://doi.org/10.1175/JCLI-D-11-00084.1.

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

Hamill, T. M., and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146**, 4079–4098, https://doi.org/10.1175/MWR-D-18-0147.1.

——, D. R. Stovern, and L. L. Smith, 2023: Improving national blend of models probabilistic precipitation forecasts using long time series of reforecasts and precipitation reanalyses. Part I: Methods. *Mon. Wea. Rev.*, **151**, 1521–1534, https://doi.org/10.1175/MWR-D-22-0308.1.

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.

Hersbach, H., and Coauthors, 2018a: ERA5 hourly data on pressure levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed 3 February 2023, https://doi.org/10.24381/cds.bd0915c6.

——, and Coauthors, 2018b: ERA5 hourly data on single levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed 3 February 2023, https://doi.org/10.24381/cds.adbb2d47.

——, and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hoopes, C. A., C. L. Castro, A. Behrangi, M. R. Ehsani, and P. Broxton, 2023: Improving prediction of mountain snowfall in the southwestern United States using machine learning methods. *Meteor. Appl.*, **30**, e2153, https://doi.org/10.1002/met.2153.

Jag, J., 2023: Avalanche chaos causes Cottonwood Canyon closures, strands thousands. *Salt Lake Tribune*, 5 April, https://www.sltrib.com/sports/2023/04/05/avalanche-chaos-causes-cottonwood/.

Judson, A., and N. Doesken, 2000: Density of freshly fallen snow in the central Rocky Mountains. *Bull. Amer. Meteor. Soc.*, **81**, 1577–1588, https://doi.org/10.1175/1520-0477(2000)081<1577:DOFFSI>2.3.CO;2.

Knowles, N., M. D. Dettinger, and D. R. Cayan, 2006: Trends in snowfall versus rainfall in the western United States. *J. Climate*, **19**, 4545–4559, https://doi.org/10.1175/JCLI3850.1.

Lackmann, G. M., and J. R. Gyakum, 1999: Heavy cold-season precipitation in the northwestern United States: Synoptic climatology and an analysis of the flood of 17–18 January 1986. *Wea. Forecasting*, **14**, 687–700, https://doi.org/10.1175/1520-0434(1999)014<0687:HCSPIT>2.0.CO;2.

Leone, G., D. Dobb, and D. Rudack, 2023: Cobb melting SLR method. National Weather Service Meteorological Development Laboratory, 27 pp., https://vlab.noaa.gov/documents/6609493/7858320/Cobb+Melting+SLR+Method.pdf.

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.

Mason, I., 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.

Mesinger, F., and Coauthors, 2006: North American regional reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, https://doi.org/10.1175/BAMS-87-3-343.

Milbrandt, J. A., A. Glazer, and D. Jacob, 2012: Predicting the snow-to-liquid ratio of surface precipitation using a bulk microphysics scheme. *Mon. Wea. Rev.*, **140**, 2461–2476, https://doi.org/10.1175/MWR-D-11-00286.1.

Mott, R., D. Scipión, M. Schneebeli, N. Dawes, A. Berne, and M. Lehning, 2014: Orographic effects on snow deposition patterns in mountainous terrain. *J. Geophys. Res. Atmos.*, **119**, 1419–1439, https://doi.org/10.1002/2013JD019880.

Mueller, M., 2001: Snow stability trends at Wolf Creek pass, Colorado. *Proc. 2000 Int. Snow Science Workshop*, Big Sky,

MT, Montana State University, 147–152, https://arc.lib.montana.edu/snow-science/item/721.

Nalli, B., and M. Mckee, 2018: How little cottonwood canyon got this way and what can be done to fix it. *Proc. Int. Snow Science Workshop 2018*, Innsbruck, Austria, Montana State University, 246–250, https://arc.lib.montana.edu/snow-science/objects/ISSW2018_O03.9.pdf.

NCEI, 2023: U.S. Climate Normals. National Centers for Environmental Information (NCEI), accessed 23 August 2023, https://www.ncei.noaa.gov/products/land-based-station/us-climate-normals.

NCEP, 2019: Upgrade NCEP Global Forecast Systems (GFS) to v15.1: Effective June 12, 2019. NCEP Service Change Notice 19-40, 8 pp., https://www.weather.gov/media/notification/scn19-40gfs_v15_1.pdf.

——, 2020: Upgrade to the RAP and HRRR analysis and forecast system, including change to location of North America Rapid Refresh Ensemble (NARRE) data: Effective December 2, 2020. NCEP Service Change Notice 20-46, 8 pp., https://www.weather.gov/media/notification/pdf2/scn20-46rap_v5_hrrr_v4_aab.pdf.

——, 2021: Upgrade NCEP Global Forecast Systems (GFS) to v16: Effective March 22, 2021. NCEP Service Change Notice 21-20, 10 pp., https://www.weather.gov/media/notification/pdf2/scn21-20gfs_v16.0_aac.pdf.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Pomeroy, J. W., and E. Brun, 2001: Physical properties of snow. *Snow Ecology*, H. G. Jones et al., Eds., Cambridge University Press, 45–126.

Potter, J. G., 1965: Water content of freshly fallen snow. CIR-4232, TEC-569, 12 pp.

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, https://doi.org/10.1029/2006GL026689.

Reeves, H. D., A. V. Ryzhkov, and J. Krause, 2016: Discrimination between winter precipitation types based on spectral-bin microphysical modeling. *J. Appl. Meteor. Climatol.*, **55**, 1747–1761, https://doi.org/10.1175/JAMC-D-16-0044.1.

Riley, C., S. Rupper, J. W. Steenburgh, C. Strong, A. K. Kochanski, and S. Wolvin, 2021: Characteristics of historical precipitation in high mountain Asia based on a 15-year high resolution dynamical downscaling. *Atmosphere*, **12**, 355, https://doi.org/10.3390/atmos12030355.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

——, S. L. Bruening, D. M. Schultz, and J. V. Cortinas Jr., 2003: Improving snowfall forecasting by diagnosing snow density. *Wea. Forecasting*, **18**, 264–287, https://doi.org/10.1175/1520-0434(2003)018<0264:ISFBDS>2.0.CO;2.

——, M. R. Butt, S. J. Reinke, and T. J. Grafenauer, 2007: Real-time forecasting of snowfall using a neural network. *Wea. Forecasting*, **22**, 676–684, https://doi.org/10.1175/WAF1000.1.

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

Schaerer, P., 1989: The avalanche-hazard index. *Ann. Glaciol.*, **13**, 241–247, https://doi.org/10.3189/S0260305500007977.

Schweizer, J., and B. Reuter, 2015: A new index combining weak layer and slab properties for snow instability prediction. *Nat. Hazards Earth Syst. Sci.*, **15**, 109–118, https://doi.org/10.5194/nhess-15-109-2015.

——, J. Bruce Jamieson, and M. Schneebeli, 2003: Snow avalanche formation. *Rev. Geophys.*, **41**, 1016, https://doi.org/10.1029/2002RG000123.

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

Serreze, M. C., M. P. Clark, R. L. Armstrong, D. A. McGinnis, and R. S. Pulwarty, 1999: Characteristics of the western United States snowpack from snowpack telemetry (SNOTEL) data. *Water Resour. Res.*, **35**, 2145–2160, https://doi.org/10.1029/1999WR900090.

Sha, Y., D. J. Gagne II, G. West, and R. Stull, 2020: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: Daily precipitation. *J. Appl. Meteor. Climatol.*, **59**, 2075–2092, https://doi.org/10.1175/JAMC-D-20-0058.1.

Steenburgh, W., 2023: *Secrets of the Greatest Snow on Earth*. 2nd ed. Utah State University Press, 226 pp.

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

Stovern, D. R., T. M. Hamill, and L. L. Smith, 2023: Improving national blend of models probabilistic precipitation forecasts using long time series of reforecasts and precipitation reanalyses. Part II: Results. *Mon. Wea. Rev.*, **151**, 1535–1550, https://doi.org/10.1175/MWR-D-22-0310.1.

The BUFKIT Warehouse, 2023: Global Bufkit profile selection. Accessed 14 September 2023, https://meteor.geol.iastate.edu/~ckarsten/bufkit/data/.

The COMET Program, 2023: Understanding NBM v4.0 snowfall products. The University Corporation for Atmospheric Research, 17 April 2024, https://www.meted.ucar.edu/education_training/lesson/10166.

Utah Department of Transportation, 2022: Traffic and transportation. Little Cottonwood Canyon environmental impact statement S.R. 210, 28 pp., https://littlecottonwoodeis.udot.utah.gov/wp-content/uploads/2022/08/LCC_FEIS_07_Traffic_Transportation.pdf.

Velasquez, P., M. Messmer, and C. C. Raible, 2020: A new bias-correction method for precipitation over complex terrain suitable for different climate states: A case study using WRF (version 3.8.1). *Geosci. Model Dev.*, **13**, 5007–5027, https://doi.org/10.5194/gmd-13-5007-2020.

Ware, E. C., D. M. Schultz, H. E. Brooks, P. J. Roebber, and S. L. Bruening, 2006: Improving snowfall forecasting by accounting for the climatological variability of snow density. *Wea. Forecasting*, **21**, 94–103, https://doi.org/10.1175/WAF903.1.

Wasserstein, M. L., and J. Steenburgh, 2023: Alta-Collins snow and liquid precipitation equivalent observations 2000–2023. Accessed 1 September 2023, https://toi.lib.utah.edu/resolve/10.7278/S50d-nsy5-8bje.

——, and W. J. Steenburgh, 2024: Diverse characteristics of extreme orographic snowfall events in Little Cottonwood Canyon, Utah. *Mon. Wea. Rev.*, **152**, 945–966, https://doi.org/10.1175/MWR-D-23-0206.1.

White, A. B., D. J. Gottas, A. F. Henkel, P. J. Neiman, F. M. Ralph, and S. I. Gutman, 2010: Developing a performance measure for snow-level forecasts. *J. Hydrometeor.*, **11**, 739–753, https://doi.org/10.1175/2009JHM1181.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 704 pp.