Evaluating Operational and Experimental HRRR Model Forecasts of Atmospheric River Events in California

Jason M. English,[a,b] David D. Turner,[a] Trevor I. Alcott,[a] William R. Moninger,[a,b]
Janice L. Bytheway,[b,c] Robert Cifelli,[c] and Melinda Marquis[a]

[a] *NOAA/Global Systems Laboratory, Boulder, Colorado*
[b] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[c] *NOAA/Physical Sciences Laboratory, Boulder, Colorado*

(Manuscript received 21 May 2021, in final form 18 August 2021)

ABSTRACT: Improved forecasts of atmospheric river (AR) events, which provide up to half the annual precipitation in California, may reduce impacts to water supply, lives, and property. We evaluate quantitative precipitation forecasts (QPF) from the High-Resolution Rapid Refresh model version 3 (HRRRv3) and version 4 (HRRRv4) for five AR events that occurred in February–March 2019 and compare them to quantitative precipitation estimates (QPE) from Stage IV and Mesonet products. Both HRRR versions forecast spatial patterns of precipitation reasonably well, but are drier than QPE products in the Bay Area and wetter in the Sierra Nevada range. The HRRR dry bias in the Bay Area may be related to biases in the model temperature profile, while integrated water vapor (IWV), wind speed, and wind direction compare reasonably well. In the Sierra Nevada range, QPE and QPF agree well at temperatures above freezing. Below freezing, the discrepancies are due in part to errors in the QPE products, which are known to underestimate frozen precipitation in mountainous terrain. HRRR frozen QPF accuracy is difficult to quantify, but the model does have wind speed and wind direction biases near the Sierra Nevada range. HRRRv4 is overall more accurate than HRRRv3, likely due to data assimilation improvements, and possibly physics improvements. Applying a neighborhood maximum method impacted performance metrics, but did not alter general conclusions, suggesting closest gridbox evaluations may be adequate for these types of events. Improvements to QPF in the Bay Area and QPE/QPF in the Sierra Nevada range would be particularly useful to provide better understanding of AR events.

KEYWORDS: Atmosphere; Atmospheric river; Freezing precipitation; Forecast verification/skill; Forecasting; Numerical weather prediction/forecasting; Probabilistic Quantitative Precipitation Forecasting (PQPF); Cloud resolving models; Mesoscale models; Model evaluation/performance

1. Introduction

Atmospheric rivers (ARs) are narrow (300–500 km wide) regions of strong water vapor transport with a meridional component, primarily in the lowest 4 km of the atmosphere, accounting for over 90% of the total poleward water vapor transport at midlatitudes (Zhu and Newell 1998; Gimeno et al. 2014). ARs are a major source of precipitation for many coastal land areas, including the west coast of North America (Ralph et al. 2004; Konrad and Dettinger 2017; Dettinger 2013; Lavers et al. 2016). Up to half of the annual precipitation in California can come from a few AR events each winter season (Guan et al. 2010; Dettinger 2011; Gershunov et al. 2019). AR events provide valuable water to replenish reservoirs and reduce drought and wildfire risks, but also cause reservoir stress, flooding, mudslides, and loss of lives and property (Corringham et al. 2019).

The impacts from ARs can be better managed through improved quantitative precipitation estimates (QPE) and quantitative precipitation forecasts (QPF), which in turn may improve lead time and accuracy of precipitation, streamflow, and reservoir storage impacts. QPE products, which are generally derived from measurements, quantify the accumulated precipitation that has occurred, while QPF products, which are generally derived from model forecasts, predict the accumulated precipitation that will occur in the future. However, both QPE and QPF products have errors and uncertainties associated with them, particularly in regions with complex terrain. Additionally, frozen precipitation presents an additional challenge, which is not accurately measured by many QPE products, if at all. An investigation of numerous satellite, radar, and gauge-based hourly QPE products found them to disagree by up to an order of magnitude in California (Bytheway et al. 2020).

ARs are challenging to model accurately due to a complex evolution of synoptic and mesoscale meteorological features (Kingsmill et al. 2006; Ralph et al. 2010; Cannon et al. 2017, 2020) as well as limited availability of observations over the Pacific Ocean for model assimilation. Numerous global modeling studies have evaluated forecast skill of ARs along the West Coast of the United States. DeFlorio et al. (2018) found forecast skill from the European Centre for Medium-Range Weather Forecasts (ECMWF) model to vary based on season and atmosphere–ocean oscillations. Lavers et al. (2020) found the ECMWF model to be too cold throughout the troposphere, and too dry with winds and water vapor flux too weak in the lower troposphere (below 900 hPa) compared to dropsonde measurements over the Pacific Ocean. Stone et al. (2020) found significantly improved forecasts when temperature and wind data from dropsondes over the Pacific Ocean are assimilated in the Navy Global Environmental (NAVGEM) Model.

Corresponding author: Jason M. English, jason.english@noaa.gov

DOI: 10.1175/WAF-D-21-0081.1

High-resolution "convection-permitting" models may be able to improve AR forecasts through improved representation of mesoscale meteorological features, terrain, and/or planetary boundary layer processes. Gowan et al. (2018) found several high-resolution models—the High-Resolution Rapid Refresh (HRRR), the North American Model (NAM), and the NCAR Ensemble—to be more accurate than coarser operational models when comparing QPF to QPE products over the western contiguous United States (CONUS) during the cool season. Among high-resolution models, two intercomparison studies have found the HRRR to perform the best (Gowan et al. 2018; Dougherty et al. 2021, manuscript submitted to *Wea. Forecasting*, hereafter DHN). Even so, several studies have noted a HRRR QPF dry bias in the Bay Area and along the Pacific Coast (Darby et al. 2019; DHN). Darby et al. (2019) suggested that HRRR terrain resolution, representation of thermodynamics, or vertical distribution of moisture may contribute to the QPF dry bias in the Bay Area, as lower tropospheric wind speed and integrated water vapor (IWV) (which can impact QPF) compared reasonably well between 3-h forecasts from the HRRRv3 and eight Atmospheric River Observatories (AROs) in the region during the 2016/17 cool season. Jeworrek et al. (2021) investigated the impacts of changing physics, microphysics, and grid spacing specifications in the Weather Research and Forecasting (WRF) Model (on which the HRRR is based) for a year of precipitation forecasts in British Columbia and concluded that the choice of cumulus and microphysics parameterizations had the largest impact on precipitation forecasts.

Quantifying precipitation forecast skill is challenging. Contingency tables are often utilized, where the forecast is compared to the observation at every point in the verification domain during a time period and categorized into hits, misses, false alarms, and correct rejections (Jolliffe and Stephenson 2011). However, slight offsets in the timing and location of precipitation result in one hit and one miss, which is especially common when evaluating models with fine horizontal grid spacing. Several approaches have been developed to address this "double penalty" problem, including object-based methods (Davis et al. 2006) and neighborhood methods (Ebert 2008). The fractions skill score (FSS) (Roberts and Lean 2008) is a commonly used neighborhood method that compares the forecast frequency to the observed frequency of an event computed in the neighborhood. The FSS is used operationally at the Met Office to verify the high-resolution model forecasts (Mittermaier et al. 2013). However, the FSS does not consider contingency table information, and several methods have been developed to incorporate both neighborhood and contingency table information (Clark et al. 2010; Schwartz 2017; Stein and Stoop 2019).

As part of the Advanced Quantitative Precipitation Information (AQPI) project (Cifelli et al. 2018; Bytheway et al. 2020), we aim to answer three primary questions in this paper: 1) How well does the operational HRRRv3 forecast precipitation of California AR events that occurred in February and March 2019, and are there patterns of biases common to the events? 2) How do precipitation forecasts from the HRRRv3 compare to HRRRv4, and can any known model changes explain QPF differences? 3) Can we identify causes of discrepancies between the QPE and QPF products?

To answer these questions, we note the following aspects of our experimental design. To minimize the risks of making overly broad conclusions based on case-dependent results, we include five AR events in our analysis, and calculate confidence intervals. To address the known uncertainties with QPE product accuracy, we compare QPF to two different QPE products: Stage IV (Lin and Mitchell 2005; Nelson et al. 2016) and a subset of gauges from the Meteorological Assimilation Data Ingest System (MADIS), which we refer to as "Mesonet." To address the known limitations of comparing models of different horizontal grid spacings to point gauges, we evaluate using both the closest grid box and neighborhood precipitation statistical methods. To help identify possible causes of model QPF errors, we compare model output to measurements of temperature, water vapor, and wind speed/direction at numerous locations. Finally, we explore whether differences between QPF and QPE are related to the occurrence of frozen precipitation, which is a known challenge with QPE products.

2. Data and methods

a. AR events studied

We study five AR events that occurred in February and March 2019 (Table 1). This time period was chosen for the following reasons: 1) all five AR events impacted the Bay Area and mountainous regions of California; 2) HRRRv3 operational output is available, and the final version of HRRRv4 code available for use; and 3) the events occurred over a relatively short 5-week time period, making a HRRRv4 retrospective run of that length feasible. All five events share some common meteorological features typical of AR events impacting California (Fig. 1). An extratropical cyclone is present off the Pacific Coast of the northwest CONUS with its associated warm conveyor belt. This warm conveyor belt contains the AR (narrow filament of large water vapor content), and a moist low-level jet in advance of the extratropical cyclone cold-frontal boundary. This boundary generally moves west to east during the event, inducing stratiform, convectively generated, and orographically enhanced precipitation (Ralph et al. 2016; Kingsmill et al. 2006; Cannon et al. 2020). A few of the AR events studied have unique characteristics. The 2–4 February 2019 event has two consecutive low pressure systems during the time period: the first starts as a large trough on 1 February that evolves into a cutoff low off the coast of California on 2 February, which then weakens and is replaced by a surface low that forms off the coast of British Columbia on 3 February, which strengthens and moves into the region on 4 February. For the 25–27 February 2019 event, the surface low remains displaced slightly farther north than in the other AR events, impacting only the northern half of California. For the 2–4 March 2019 event, the surface low (and corresponding cold front) is weaker than in the other AR events.

TABLE 1. Summary of AR events studied.

Event	Period averaged	6-h Stage IV comparisons	6-h Mesonet comparisons	1-h Mesonet comparisons
2–4 Feb 2019	1200 UTC 2 Feb to 1200 UTC 4 Feb	6-h cadence: 9 valid times total	6-h cadence: 9 valid times total	3-h cadence: 18 valid times total
13–15 Feb 2019	0600 UTC 13 Feb to 0600 UTC 15 Feb	6-h cadence: 9 valid times total	6-h cadence: 9 valid times total	3-h cadence: 18 valid times total
25–27 Feb 2019	0600 UTC 25 Feb to 0600 UTC 27 Feb	6-h cadence: 9 valid times total	6-h cadence: 9 valid times total	3-h cadence: 18 valid times total
2–3 Mar 2019	0600 UTC 2 Mar to 0600 UTC 4 Mar	6-h cadence: 9 valid times total	6-h cadence: 9 valid times total	3-h cadence: 18 valid times total
5–6 Mar 2019	0600 UTC 5 Mar to 0000 UTC 7 Mar	6-h cadence: 7 valid times total	6-h cadence: 7 valid times total	3-h cadence: 14 valid times total

The duration of the AR events is approximately two to five days each, based on accumulated precipitation in the AQPI domain (Table 1). For this study, we average precipitation across the 48-h time period in which most of the accumulated precipitation occurred in the AQPI domain (Fig. 2) based on visual inspection of radar reflectivity maps.

b. RAP/HRRR model

The RAP/HRRR is an hourly updated data assimilation and numerical weather prediction system, run operationally at NOAA/National Centers for Environmental Prediction (NCEP) (Benjamin et al. 2016; Alexander et al. 2017). The RAP encompasses a region significantly larger than North America at 13-km horizontal grid spacing, while the HRRR encompasses slightly more than the contiguous United States at 3-km grid spacing (Fig. 2a). The RAP and HRRR both have 51 vertical levels and are initialized every hour, assimilating a variety of satellite- and surface-based datasets (Benjamin et al. 2016). The RAP provides initial and lateral boundary conditions for the HRRR, although in the latest version (RAPv5/HRRRv4), the HRRR receives its initial conditions through a 36-member HRRR ensemble analysis system instead of the RAP. As with any typical operational system, both the RAP and the HRRR undergo continuous

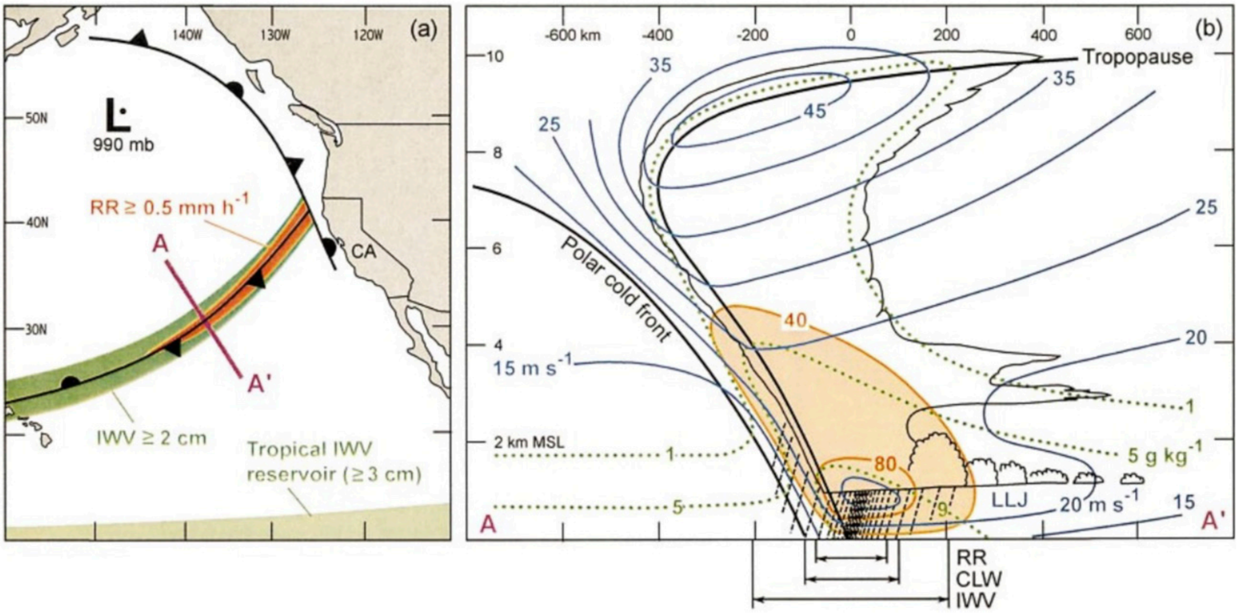

FIG. 1. Conceptual representation of a typical atmospheric river over the northeastern Pacific Ocean. (a) Plan-view schematic of concentrated IWV (IWV ≥ 2 cm; dark green) and associated rain-rate enhancement (RR ≥ 0.5 mm h^{-1}; red) along a polar cold front. The tropical IWV reservoir (0.3 cm; light green) is also shown. The bold line AA′ is a cross-section projection for (b). (b) Cross-section schematic through an atmospheric river [along AA′ in (a)] highlighting the vertical structure of the alongfront isotachs (blue contours; m s^{-1}), water vapor specific humidity (dotted green contours; g kg^{-1}), and horizontal alongfront moisture flux (red contours and shading; $\times 10^5$ kg s^{-1}). Schematic clouds and precipitation are also shown, as are the locations of the mean width scales of the 75% cumulative fraction of perturbation IWV (widest), cloud liquid water (CLW), and RR (narrowest) across the 1500-km cross-section baseline (bottom) (from Ralph et al. 2004).

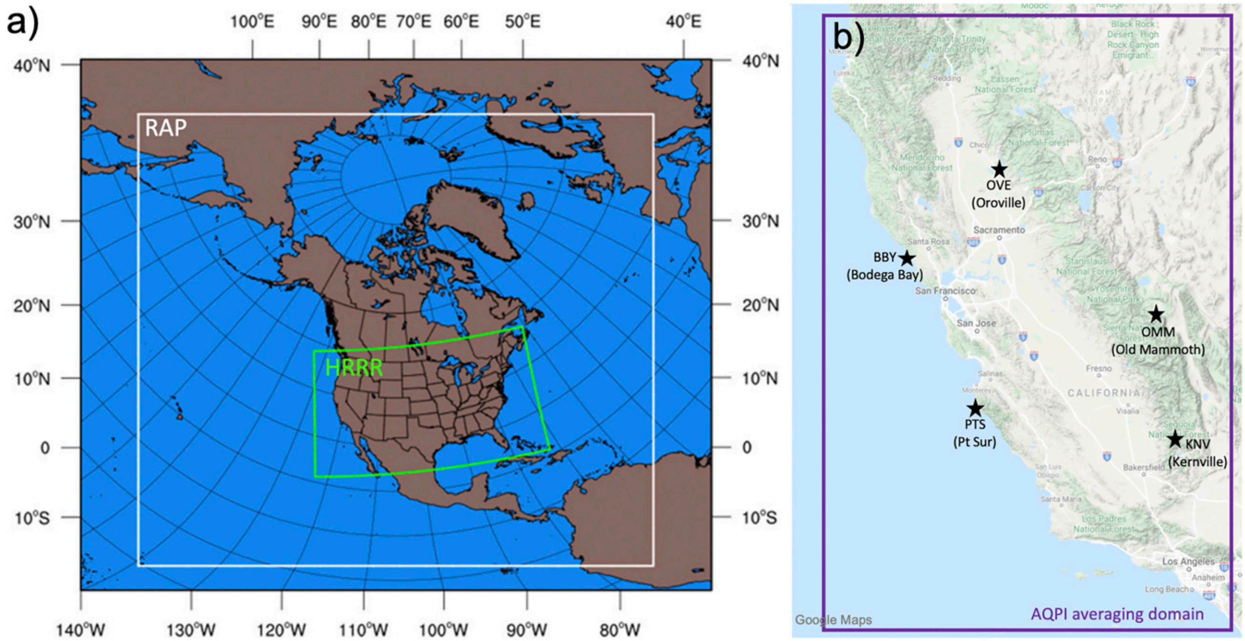

FIG. 2. (a) RAP/HRRR domain. (b) Evaluation domain and ARO station locations (stars).

development with new versions introduced into operations approximately every 2 years, which provides benefit via improved forecasts over time, but also creates challenges with model assessment, as model versions regularly change. The latest version, RAPv5/HRRRv4, became operational on 2 December 2020. RAPv5/HRRRv4 includes many assimilation and physics improvements over the previous version (RAPv4/HRRR3), including planetary boundary layer (PBL) code updates, enhanced gravity wave drag, assimilation of *GOES-16* radiances, a 36-member HRRR 3-km ensemble for the DA, improved hydrometeor assimilation, and some new observations and data assimilation methods (Benjamin et al. 2016; Alexander et al. 2017).

For each AR event, we evaluate output from the operational HRRR (HRRRv3) and the experimental HRRR (HRRRv4) corresponding to the dates of the events. HRRRv3 output was downloaded from existing operational runs, while HRRRv4 output was produced by running retrospective simulations at the NOAA Global Systems Laboratory in 2019 (see approach in James and Benjamin 2017). For this study we used frozen HRRRv4 code, which became the operational RAP/HRRR system on 2 December 2020. For these retrospective simulations, HRRRv4 was run with 36-h forecasts twice a day at 0000 and 1200 UTC, 18-h forecasts at 0300, 0600, 0900, 1500, 1800, and 2100 UTC, and 3-h forecasts at the remaining hourly times. The number of forecasts utilized for each AR event is summarized in Table 1.

c. *Stage IV and Mesonet QPE products*

We utilize two QPE products for this work: NCEP Stage IV (Lin and Mitchell 2005), and "Mesonet" (a gauge network from NOAA's MADIS program). The two QPE products are not completely independent from one another, as some gauges from Mesonet are utilized to produce the Stage IV product.

Stage IV is a regional hourly/6-hourly multisensor (radar plus gauges) product produced by the twelve River Forecast Centers (RFCs), and mosaicked into a 4.7-km-grid national product at NCEP. The 6-hourly Stage IV is reprocessed based on new information for a time period of up to 7 days after the valid time. Each RFC conducts some manual quality control, and retains authority over which specific datasets contribute to its respective analysis. The California–Nevada RFC (CNRFC), releases a 6-hourly product based solely on gauges and climatology (i.e., without using radar-derived precipitation estimates). CNRFC starts with the mountain mapper (Schaake et al. 2004; Zhang et al. 2011), which is an orography-adjusted gauge interpolation product based on the Parameter-elevation Regressions on Independent Slopes Model (PRISM) climatology (Daly et al. 2008). CNRFC adjusts the mountain mapper grid with measurements from 600 to 800 trusted gauges, including county alert gauges, out of 2000 total available in the CNRFC domain in 2019 (R. Hartmann 2020, personal communication). CNRFC also uses a "1.2 snow correction factor" (i.e., gauge-reported accumulations are increased by 20% when the precipitation type is determined to be snow) because of known persistent undercatch by many gauges. More details on Stage-IV are provided in Nelson et al. (2016). For our work, the Stage IV grid is remapped to a 3-km grid using an equally weighted mean of nearest-neighbor and budget interpolation methods to directly compare to the HRRR.

The "Mesonet" data used is from three gauge networks provided from MADIS: Remote Automated Weather Stations (RAWS) (https://www.nifc.gov/aboutNIFC/about_RAWS.html), Hydrometeorological Automated Data System (HADS) (Kim et al. 2009), and MesoWest (https://mesowest.utah.edu/). These are all national surface networks and report liquid precipitation though MADIS; snow and ice measurements were not used in calculating the liquid precipitation totals (G. Pratt, MADIS Lead for NOAA Research, 2021, personal communication). Both HADS and MesoWest provide data from several subnetworks through MADIS. A complete listing of the networks included in MADIS is available at https://madis.noaa.gov/mesonet_providers.shtml. We do not include any Citizen Weather Observer Program stations in the database. We utilize all stations reporting data during the time period of each AR event, requiring the 1-h accumulation to be between zero and 76 mm as a rough quality control check, which resulted in 420–480 Mesonet gauges included in each event. For closest grid box comparison, we create a 3-km grid and assign each gauge value to the grid box that encompasses the gauge location. If there is more than one gauge located in the same grid box, the gauge with the larger precipitation value is used. This was a rare occurrence, and collocated gauges reported data in the same grid box in only 3% of the Mesonet data.

d. Metrics

Since a primary goal of the AQPI project is to improve short-term monitoring of precipitation, streamflow, and coastal flooding in the San Francisco Bay Area, we compare 1 and 6-h accumulated precipitation (acc) from two QPE products (Stage IV and Mesonet) to HRRR QPF at two forecast lead times: 1 and 6 h. Due to the cadence of HRRRv4 cycling, 1-h forecasts are available every hour, while 6-h forecasts are available every 3 h. We compare average precipitation bias (mm) as well as commonly used contingency statistics at several thresholds including frequency bias, probability of detection (POD), false alarm ratio (FAR), critical success index (CSI; Gilbert 1884; Donaldson et al. 1975), and equitable threat score (ETS; Gilbert 1884; Schaefer 1990). We evaluate spatial comparisons as well as an areal average across our designated AQPI domain (33.3°–41.4°N, 118.2°–123.8°W). Precipitation is evaluated via both closest grid box and the neighborhood maximum (NM) method (Schwartz 2017). The NM method creates a neighborhood for both the forecast and the observation, and considers any precipitation that overlaps as a hit. We choose the NM method for several reasons: 1) it can easily be applied to both gridded and point (gauge) datasets; 2) it incorporates both neighborhood and contingency table information, 3) it was found to be the most realistic of three neighborhood methods evaluated (Schwartz 2017); and 4) we wanted to explore its usefulness for QPF/QPE in mountainous terrain.

To better understand model performance, we also compare forecasts of temperature, wind speed/direction, and integrated water vapor to meteorological terminal aviation routine weather report (METAR) (Turner et al. 2020) and ARO measurements. A map of the domain over which we average our fields as well as ARO station locations is provided in Fig. 2b.

3. Results

a. Spatial patterns of precipitation

Mean 6-hourly QPF and QPE for each of the five AR events are computed for the time periods during the 48-h period in which most of the precipitation occurred (Table 1). Spatial maps of mean 6-h Stage-IV QPE (Fig. 3) show some common features: All AR events have more precipitation at higher elevations (particularly the coastal and Sierra Nevada mountains), and less precipitation at lower elevations (particularly the Central Valley). This well-known precipitation dichotomy reflects the influence of orographic lifting and precipitation enhancement in both the coastal mountains and the Sierra Nevada range. Precipitation in the Bay Area varies for each event, but is usually substantial [>10 mm $(6\,\mathrm{h})^{-1}$], especially in the Santa Cruz mountains south of San Francisco Bay and the coastal mountains north of San Francisco. Compared to the other events, precipitation was relatively higher for the 13–15 February 2019 and 25–27 February 2019 AR events, with more precipitation in the northern half of the AQPI domain. Bias maps of HRRRv3 and HRRRv4 show a reasonable visual agreement with Stage IV spatial gradients of precipitation, but often are wetter in the Sierra Nevada range and drier along the Pacific Coast, particularly in the Bay Area. One exception is the 2–4 March 2019 event, where HRRR did not have a clear bias in the Sierra Nevada range, although total precipitation was less for this event in general. This case highlights the variation that individual events can have on model forecast accuracy.

Spatial patterns of precipitation from Mesonet QPE (Fig. 4) appear similar to that of Stage IV (Fig. 3), although the two QPE products were not directly compared. Both HRRRv3 and HRRRv4 compare reasonably well to Mesonet, but as with Stage IV, generally are wetter in the Sierra Nevada range and drier in the Bay Area. This inferred dry bias in the Bay Area (assuming the QPE products are correct) is also noted in other modeling studies (Darby et al. 2019; DHN).

Bias maps of HRRRv3 and HRRRv4 suggest small differences between the two model versions (Figs. 3 and 4). HRRRv4 usually has smaller biases than HRRRv3 compared to Mesonet, as shown by the number of data points within each bias range (see Fig. 4 legend). HRRRv4 has more data points within 3 mm of the gauge measurements than HRRRv3 for four out of five AR events, and has the same number for the remaining event, suggesting the newer version of the model improves forecasts of AR events. When summing across all five AR events, a larger percentage of HRRRv4 data points are within 3 mm of the gauge values than HRRRv3 (75% versus 72%, respectively). This improvement comes from both a reduction in the dry bias (four out of five events) and wet bias (three out of five events).

b. Categorizing precipitation patterns and biases

To more clearly quantify patterns of precipitation averages and biases, we average all five AR events across the AQPI domain and calculate mean 6-h precipitation for three altitude ranges: 0–4200 m (entire domain), 0–1000 m (mostly outside the Sierra Nevada range), and 1000–4200 m (mostly the Sierra

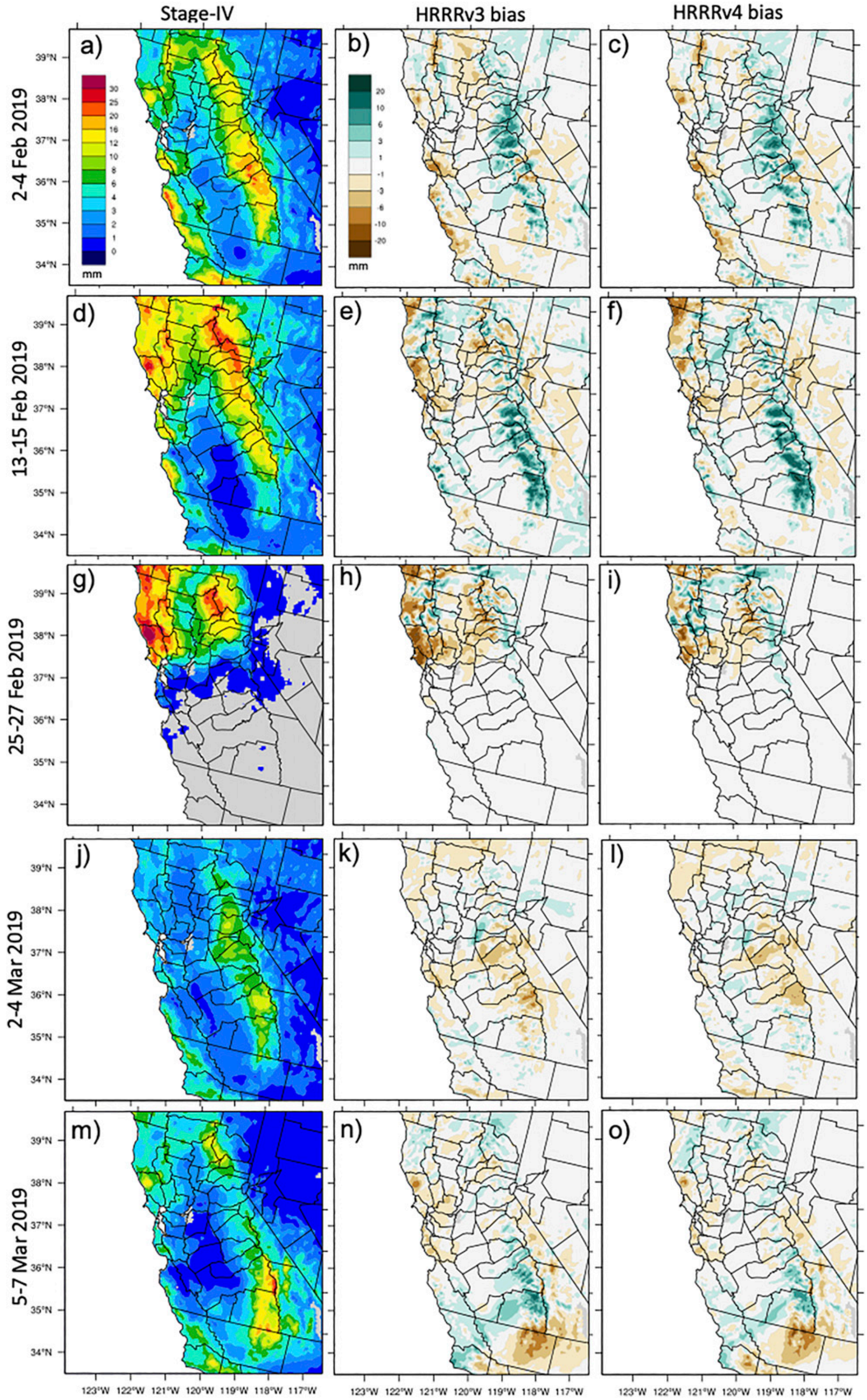

FIG. 3. Average 6-h accumulation (mm) for Stage IV, HRRRv3 bias (6-h lead time), and HRRRv4 bias (6-h lead time), averaged across the peak 48-h time period of each event (Table 1). HRRR bias is calculated as HRRR − Stage IV; blue–green colors are a model wet bias and brown colors are a dry bias.

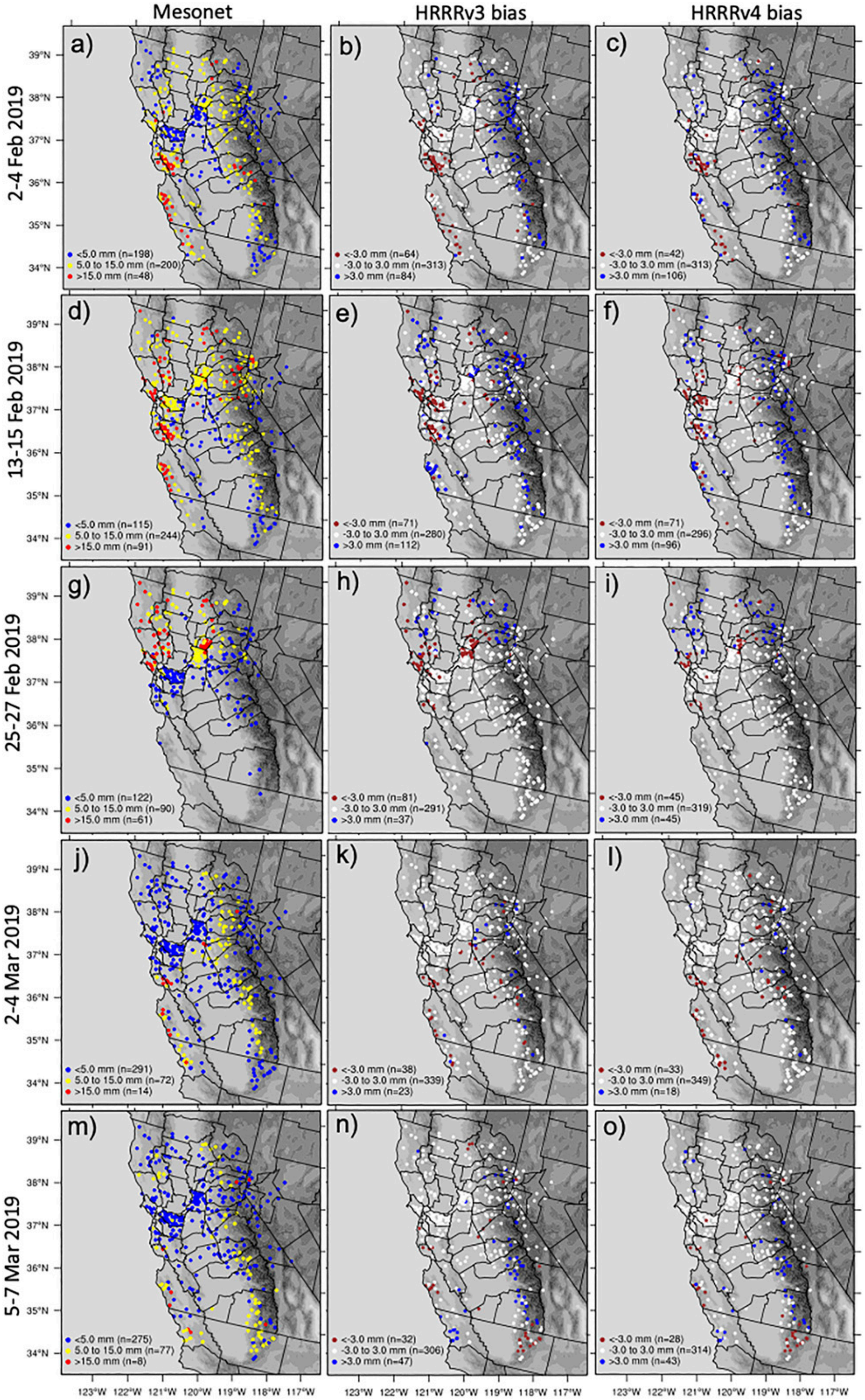

FIG. 4. Average 6-h accumulation (mm) for (left) Mesonet, (center) HRRRv3 bias (6-h lead time), and (right) HRRRv4 bias (6-h lead time), averaged across the peak 48-h time period of each event (Table 1). HRRR bias is calculated as HRRR − Mesonet; blue colors are a model wet bias, and red colors are a dry bias.

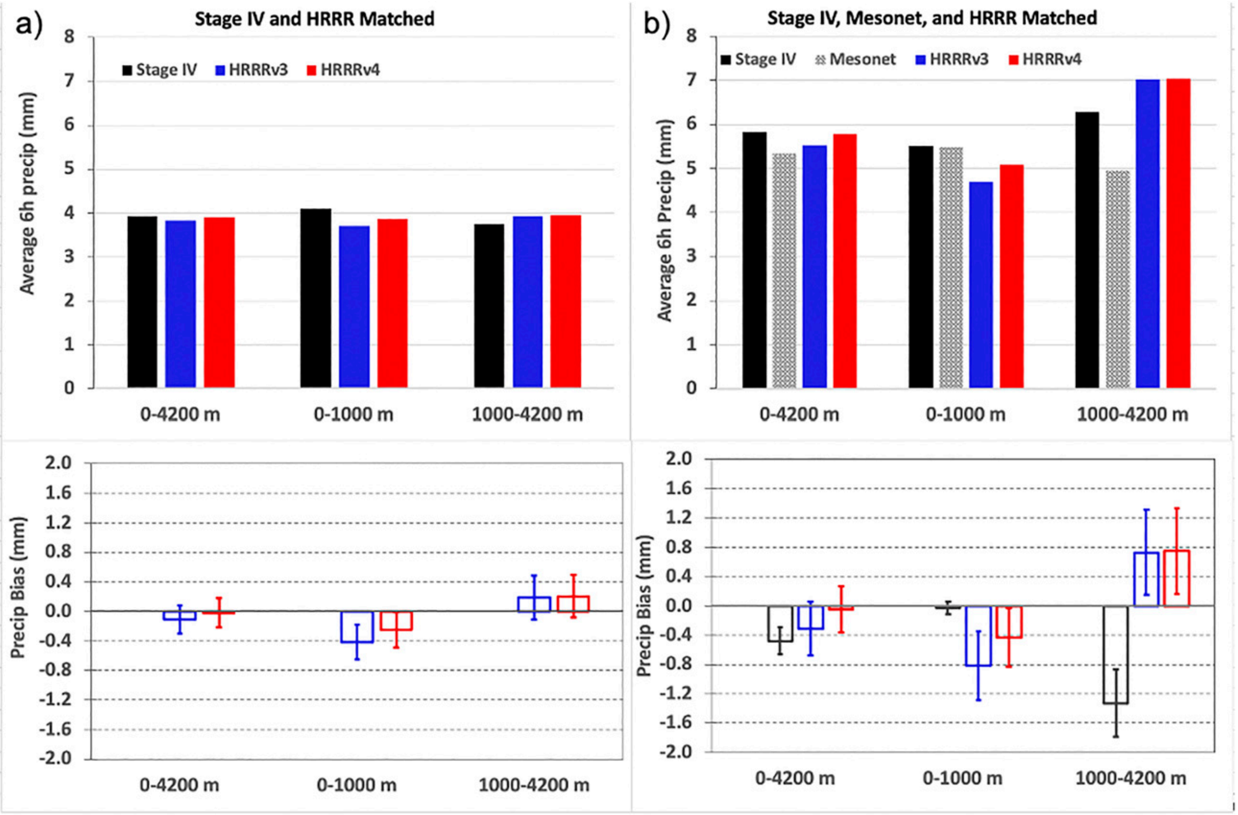

FIG. 5. Average 6-h accumulation (mm) for Stage IV, Mesonet, and HRRR (6-h lead times), averaged across the AQPI domain for all five AR events, at three different altitude ranges. (a) Stage IV and HRRR matched comparisons (grid boxes common to Stage IV and HRRR data included.) (b) Stage IV, Mesonet, and HRRR matched comparisons (grid boxes common to Stage IV, Mesonet, and HRRR data included.) Precipitation bias is relative to Stage IV, and error bars represent 95% confidence interval of the standard error of the mean.

Nevada range). Two sets of matching comparisons are conducted: a comparison between Stage IV and HRRR (Fig. 5a), and a comparison between Stage IV, Mesonet, and HRRR (Fig. 5b). In both panels, bias is calculated with respect to Stage IV. When considering only points and times with Mesonet availability (Fig. 5b), Stage IV and HRRR precipitation are roughly 20%–30% larger than when considering all grid points in the domain (Fig. 5a). Recall that Stage IV is a continuous, gridded product available across the entire AQPI domain, while Mesonet consists of 420–480 individual gauge locations. This suggests that locations with Mesonet gauges usually receive more precipitation than an average location within the AQPI domain. Comparing Stage IV to Mesonet (averaging QPE in grid boxes that are common to both products), the two QPE products have excellent agreement at 0–1000 m, but Mesonet is significantly drier than Stage IV at 1000–4200 m (Fig. 5b). This difference is primarily attributed to the presence of frozen precipitation at higher elevations, which is included in the Stage IV product but not Mesonet. The influence of frozen precipitation is discussed further in section 3c.

When averaging across all altitudes (0–4200 m), average accumulated precipitation for both HRRRv3 and HRRRv4 are in the range of Stage IV and Mesonet QPE. However, both HRRR versions are significantly drier than both Stage IV and Mesonet at 0–1000 m, and are wetter than Stage IV and Mesonet at 1000–4200 m. These opposing biases as a function of altitude are discussed further in the following sections.

Categorical precipitation forecast performance is quantified at 0.25-, 2.5-, and 10-mm thresholds (6-h accumulation) for several metrics, including frequency bias and CSI. ETS was also calculated, and while the values were smaller than CSI (due to random hits included in the ETS calculation), trends and conclusions were similar (not shown). We also calculate standard error 95% confidence intervals via the statistical bootstrapping technique of Hamill (1999) using 1000 permutations, with replacement. Frequency bias is the ratio of forecast and observation frequency counts; hence a ratio equal to one represents a perfect forecast. Frequency bias for both HRRR versions relative to Stage IV ranges from 0.89 to 0.96 at all three thresholds over all three altitude ranges (Fig. 6), indicating the HRRR precipitation frequency at each of the thresholds is slightly lower than Stage IV. Compared to Mesonet, frequency bias is also low over the 0–1000-m altitude range for both HRRR versions, while frequency bias is high over the 1000–4200-m altitude range. As noted above, the discrepancy between the HRRR versions and Mesonet over

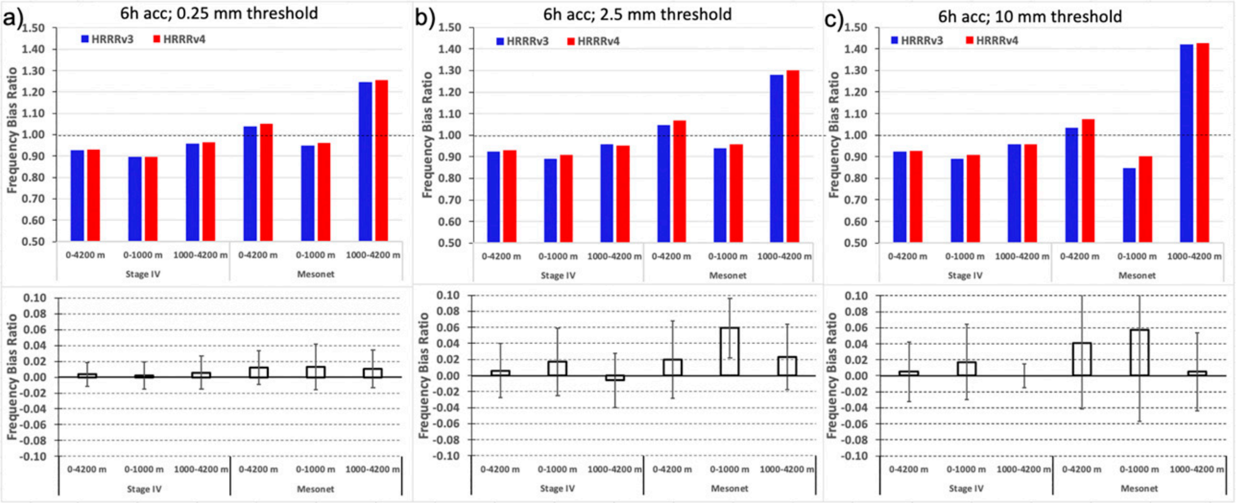

FIG. 6. Frequency bias of 6-h accumulation (mm) for HRRR (6-h lead times) against Stage IV and Mesonet, averaged across the AQPI domain for all five AR events, at three different altitude ranges, and three thresholds: (a) 0.25, (b) 2.5, and (c) 10 mm. Bar heights in the bottom panels are the difference in frequency bias between HRRRv3 and HRRRv4, with error bars representing the 95% confidence interval for these differences based on the bootstrapping technique (Hamill 1999).

the 1000–4200-m altitude range is likely due to the presence of frozen precipitation which is not included in the Mesonet database. HRRRv4 usually has a higher frequency of exceeding each precipitation threshold than HRRRv3 (in 16 out of 18 of the comparisons in Fig. 6); however, most of these differences are not significant at the 95% confidence level. The one exception is with Mesonet over the 0–1000-m altitude range at 2.5-mm threshold, where HRRRv4 has significantly higher frequency bias than HRRRv3, which translates to a forecast improvement since the HRRRv4 frequency bias is closer to one.

CSI for both HRRR versions is generally larger (better) at lower thresholds than at higher thresholds compared to both Stage IV and Mesonet (Fig. 7). This is a common occurrence, as high-intensity/low-frequency events are typically more challenging for models to forecast accurately. At the 0.25- and 2.5-mm thresholds, CSI for both HRRR versions is larger (better) over the 0–1000-m altitude range than 1000–4200 m, while at the 10-mm threshold, CSI for both HRRR versions is larger (better) over the 1000–4200-m altitude range. CSI for HRRRv4 is usually larger than HRRRv3 compared to both Stage IV and Mesonet across all three altitude ranges, particularly at higher thresholds, suggesting the newer model version improves precipitation forecasts. These improvements are statistically significant at 95% confidence versus Stage IV at all thresholds, and versus Mesonet at larger thresholds (Fig. 7).

Sometimes higher CSI may arise from higher frequency bias, producing seemingly improved forecast skill due to excessive frequency of precipitation occurrence (Baldwin and Kain 2006). In an effort to further explore the reasons for improved CSI for HRRRv4 over HRRRv3, we expand our contingency table evaluation and provide performance diagrams containing POD, success rate, frequency bias, and CSI based on the work of Roebber (2009) (Fig. 8). HRRRv4 has improved forecasts over HRRRv3 for most metrics, with higher CSI, POD, and success rate. As we discussed previously, HRRRv4 also consistently has higher frequency bias than HRRRv3. This translates to a HRRRv4 improvement over HRRRv3 when frequency bias is less than one, and a degradation when frequency bias is greater than one. Comparisons to Mesonet over the 1000–4200-m range are relative outliers to the other comparisons, which again we attribute to the presence of frozen precipitation not included in Mesonet.

c. *Exploring the contributions of frozen precipitation to QPE/QPF discrepancies*

It is challenging to accurately assess HRRR QPF since QPE products have known biases, particularly with frozen precipitation and in mountainous terrain. As we mentioned previously, the Mesonet gauges in our study do not measure snowfall and cover only a small portion of the AQPI land area. Stage IV has known biases as well. The CNRFC manually adjusts Stage IV precipitation at high elevations due to known challenges with frozen precipitation. Several studies have explored QPE and QPF errors relevant to our work. Smalley et al. (2014) compared Stage IV precipitation to *CloudSat* and found the former often misses precipitation in high terrain and at temperatures below freezing, with Stage IV precipitation detection three times lower and accumulation 10% lower than *CloudSat* in the California–Nevada (CN) region. Lundquist et al. (2019) concluded that QPE products have so much error with high mountain precipitation that model forecasts may be a more reliable estimate in these regions than the QPE products themselves. These challenges highlight the need to improve QPE products in mountainous regions.

To help identify the contributions of frozen precipitation to differences between QPE and QPF, we quantify 6-h accumulated precipitation for three temperature regimes based on 2-m temperature from HRRRv4 at each hourly analysis time:

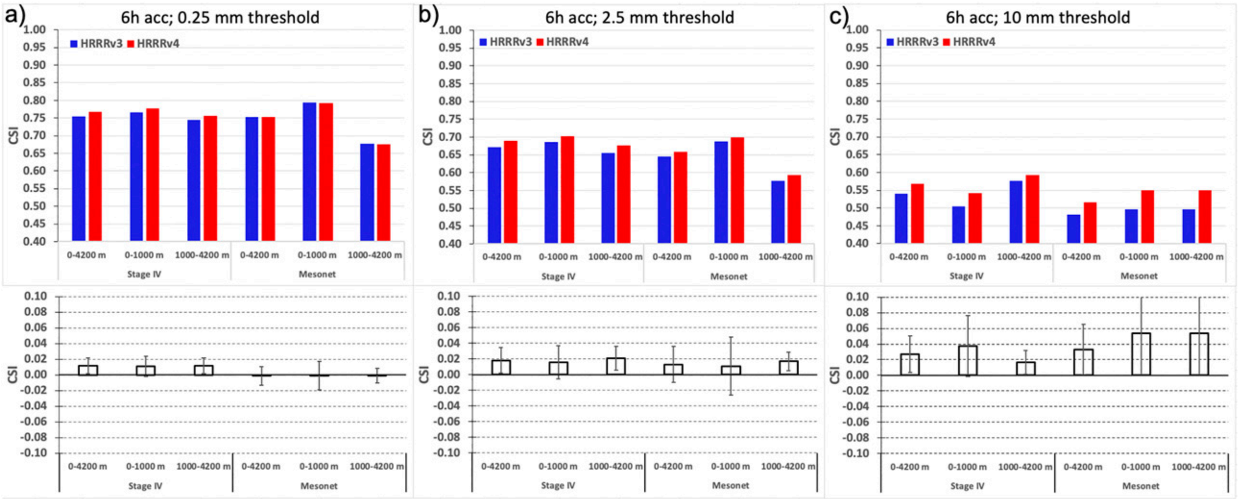

FIG. 7. As in Fig. 6, but for critical success index (CSI) instead of frequency bias.

0–400 K (i.e., all data), 273–400 K, and 0–273 K (Fig. 9a). Again, these are matched comparisons, meaning that data are averaged only when available for all four datasets; hence, all of the datasets are averaged only in grid boxes where Mesonet gauges are located and reporting data. While there is reasonable agreement between the QPE and QPF products at 0–400 and 273–400 K, there is significant disagreement below freezing. Mesonet 6-h precipitation is approximately half that reported by Stage IV and both HRRR versions at 0–273 K. It is interesting that Mesonet reports significant precipitation below 273 K, suggesting that some gauges are still measuring some precipitation at subfreezing 2-m temperatures. Additionally, as we note later, HRRRv4 has a cold bias resulting in some error in the defined boundaries of our temperature regime evaluation.

Comparisons between the four QPE/QPF products over the 1000–4200-m altitude range at four different temperature regimes provides further insight into the influence of freezing and/or frozen precipitation (Fig. 9b). Note again that due to a model cold bias, using HRRRv4 temperature to categorize liquid/frozen precipitation may have some errors. For this reason, we categorize QPE/QPF using two different temperature limits: 273 and 276 K. While there is significant disagreement between the four products across the whole temperature regime (0–400 K), it is mostly coming from temperatures below 273 K. Agreement is improved at temperatures 273–400 K, and further improved if we use a slightly warmer range of 276–400 K. In the 276–400-K temperature regime, the four products all compare very well and are not significantly different at the 95% confidence level. This result suggests the vast majority of the discrepancies occur when frozen precipitation is present. When the precipitation is liquid, HRRR compares very well to Stage IV and Mesonet in this mountainous terrain. Since all five of the AR events in this study were known to have snowfall in the Sierra Nevada range, we conclude that errors in the QPE products are a significant contribution to inferred HRRR QPF biases. It is difficult to quantify whether Stage IV or HRRR are more accurate in the 0–273-K temperature regime, however.

The first half of the 13–15 February 2019 AR event was relatively warm, with a few periods of abrupt rises in mountain snow levels (Hatchet et al. 2020). To further explore whether the discrepancies between QPF/QPE products in the Sierra Nevada range are due to errors in QPE products when it is snowing, we average QPF and QPE over two periods: a warm period (0600 UTC 13 February–1800 UTC 14 February), and a cool period (1800 UTC 14 February–0600 UTC 15 February) (not shown). All of the Sierra Nevada range in Plumas County north of Lake Tahoe (which is lower in elevation than most of the Sierra Nevada range in California) remained above 273 K during the first period, suggesting that most of the precipitation fell as rain in these locations as well. Correspondingly, HRRRv4 compared well to Stage IV and Mesonet in the first period, while in the second period, which was mainly snow, HRRRv4 was wetter. This again suggests that the presence of snow is causing QPE and QPF discrepancies, even over the same geographic region during the same AR event.

Another approach to compare QPF to QPE products is to separate model QPF into its liquid/frozen components, and compare model liquid QPF to liquid QPE products and/or model frozen QPF to snow QPE products. However, since both HRRR versions have a near-surface cold bias in the Sierra Range, the model likely partitions too much snow and not enough liquid, even if total QPF may be accurate.

d. Comparing other HRRR meteorological fields to observations

To better understand possible causes of model precipitation errors, we compare 6-h forecasts from HRRR meteorological fields to nearby METAR sites and ARO stations. Note that both METAR and ARO data are assimilated by the HRRR, so good agreement in their fields should be expected near analysis time, but model errors will likely grow at longer lead times. Low-level

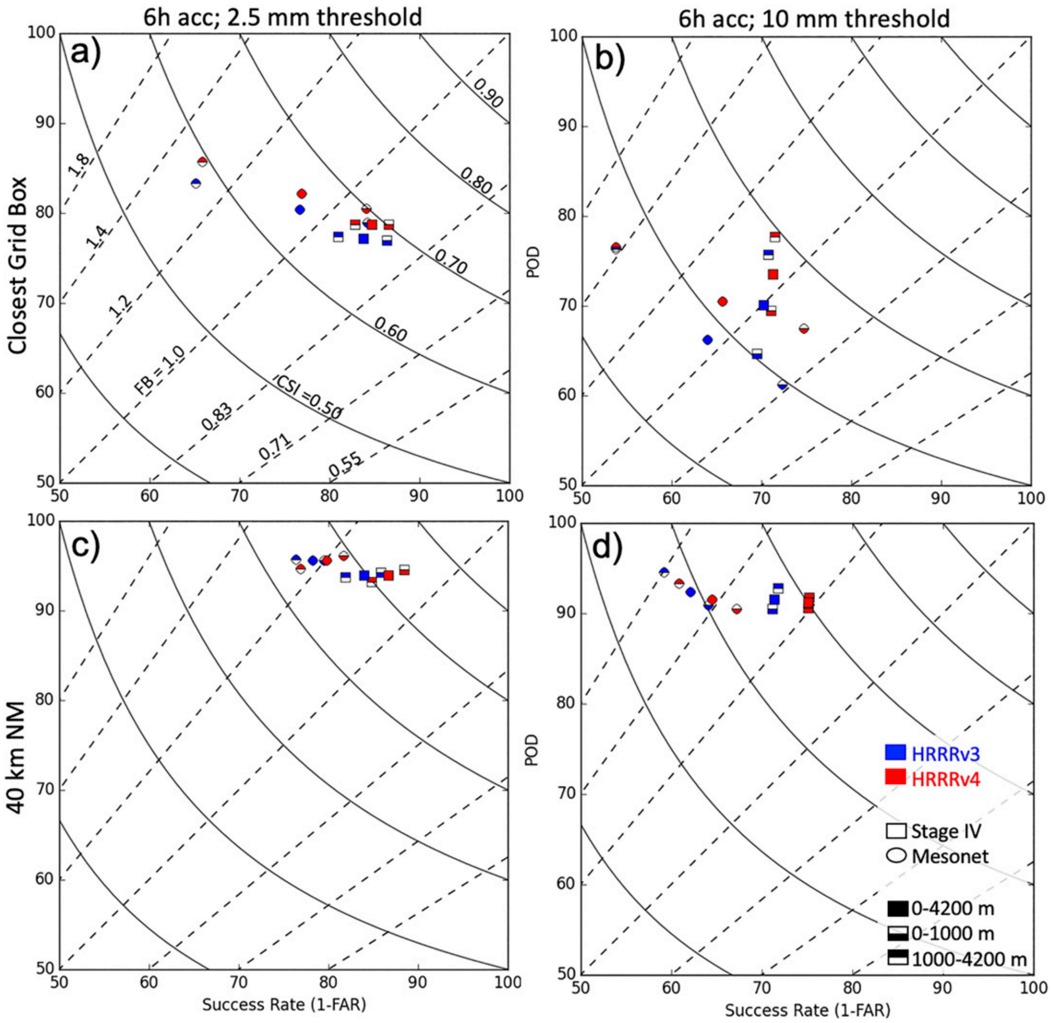

FIG. 8. Performance diagrams of CSI, frequency bias, POD, and FAR for HRRRv3 (6-h lead times; blue colors) and HRRRv4 (6-h lead times; red colors) compared to Stage IV (boxes) and Mesonet (circles) at three altitude ranges, averaged across the AQPI domain for all five ARs. (a) Closest grid box; 2.5-mm threshold. (b) Closest grid box; 10-mm threshold. (c) 40-km NM; 2.5-mm threshold. (d) 40-km NM; 10-mm threshold.

wind speed biases could represent errors in several meteorological processes related to QPF, such as errors in frontogenesis or the low-level jet ahead of the cold front of an extratropical cyclone, or errors with orographic forcing. Temperature biases could impact precipitation amount and phase. Time series of 2-m temperature and 10-m wind speed for each AR event compare reasonably well between both HRRR versions and METARs across California, but generally the model forecast temperatures are too cold and winds are too strong (Fig. 10). While both model versions forecast similar near-surface temperatures and wind speeds, HRRRv4 is typically colder with stronger winds. Both HRRR versions have a persistent wintertime diurnal temperature bias pattern in California which affects forecasts for all five AR events, with the largest bias at about 0300 UTC each day (−1.0°C cold bias), and the smallest bias at 1500 UTC each day (−0.2°C cold bias) (not shown). Looking at the spatial distribution of biases for the 13–15 February 2019 AR event, most of the cold biases are in the Central Valley (not shown), even though HRRR QPF compared favorably to Stage IV and Mesonet in that region (Figs. 3f and 4f). In the Bay Area, the HRRRv4 10-m wind speed bias varies with location, with model wind speeds too strong at some places and too weak at others (not shown). HRRRv4 has cold biases at a few locations in the Sierra Nevada range, suggesting too much model precipitation is falling as snow rather than rain; however, this does not explain the total (liquid plus frozen) precipitation wet bias over the region relative to Stage IV.

In the Bay Area, HRRR profiles of average temperature and winds compare fairly well to the ARO station at Bodega Bay (Fig. 11). Both modeled and observed temperatures decrease with height (Fig. 11a), wind speeds increase with height (Fig. 11b), and wind direction shifts from southerly to southwesterly with height (Fig. 11c). The HRRR is a little too warm at the surface and too cold aloft. The average lapse rate

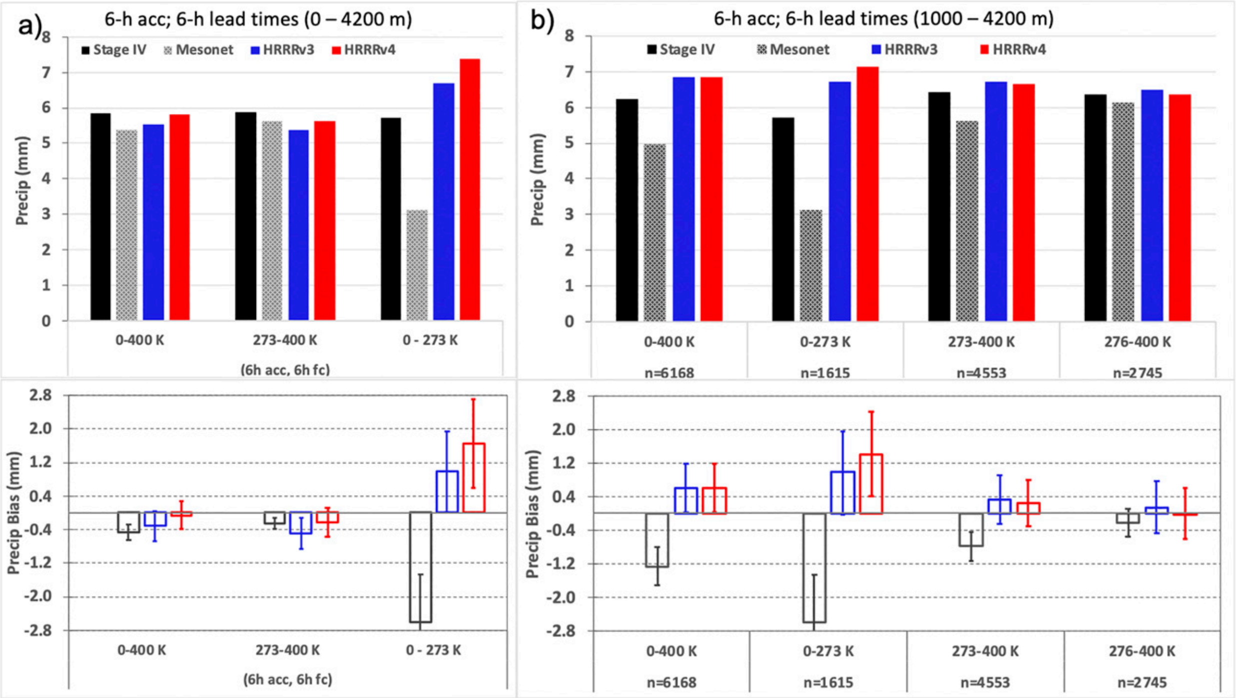

FIG. 9. Average 6-h accumulation (mm) for Stage IV, Mesonet, and HRRR (6-h lead times), averaged across the AQPI domain for all five AR events, at the temperature ranges noted: (a) 0–4200 and (b) 1000–4200 m. Output is averaged across the time period of each event, and normalized to the number of grid boxes available at each forecast output interval. Precipitation bias is relative to Stage IV, and error bars represent 95% confidence interval of the standard error of the mean.

itself does not explain the HRRR QPF dry bias in the region, but the temperature profile may be an indication of advection errors at different levels. HRRR wind speed and direction also compare relatively well to the ARO station at Bodega Bay (Figs. 11e,f). HRRRv3 and HRRRv4 mean wind speeds are slightly stronger than observed from about 0.5- to 2-km altitude. HRRRv3 mean wind direction compares well with the ARO station at Bodega Bay, while HRRRv4 is a little too southerly (clockwise). This HRRRv4 wind direction bias could produce the QPF dry bias; however, HRRRv3 also had a QPF dry bias even though wind direction compared well. Also, when looking at individual AR events, the HRRR wind speed and direction biases vary (not shown), despite a consistent HRRR dry bias in the Bay Area for all five AR events, suggesting wind biases are not the primary factor causing QPF biases. However, differences are not significant for individual events. Our comparisons of HRRR wind speed and direction to ARO stations in the Bay Area generally agree with other studies. Darby et al. (2019) found 3-h forecasts from the HRRRv3 wind speed to compare reasonably well to eight ARO stations in the Bay Area. DHN also found a small clockwise HRRR wind direction error, and suggested that may be responsible for the HRRR QPF dry bias in the Bay Area. Overall, we conclude that it is possible that errors in HRRR temperature, wind speed, or wind direction biases are indicative of possible model errors with meteorological features resulting in the HRRR QPF dry bias in the Bay Area, but these biases do not appear to play a dominant role in the QPF bias, and further work is needed to explore the QPF dry bias in the Bay Area.

Near the Sierra Nevada Range, HRRR profiles of average temperature and winds compare fairly well to the ARO station at Oroville (Fig. 12). Both modeled and observed temperatures decrease with height (Fig. 12a), wind speeds increase with height (Fig. 12b), and wind direction shifts from southeasterly to southwesterly with height (Fig. 12c). HRRR temperatures near the surface and aloft are a little colder than observed at the ARO station in Oroville (Fig. 12d). However, the lapse rate is not significantly different than observed, and the cold biases likely would impact precipitation phase more than total precipitation, which does not explain the total QPF wet bias in the HRRR. Near the surface, HRRRv4's cold bias is more pronounced than HRRRv3. This cold bias of approximately 1.5°C could be contributing to some errors in the frozen precipitation analysis discussed in section 3c, which is a reason why we constrained the "liquid" temperature regime to 276–400 K. At Oroville, the HRRR wind speed is a little too strong in the lowest 1 km of the atmosphere, and the HRRR wind direction is too southerly by about 10–20° in the lowest 2.5 km of the atmosphere (Figs. 12e,f), which may contribute to QPF errors if the wind direction error is large enough to affect the magnitude of orographic forcing for upward motion. Neiman et al. (2013) concluded that heavy precipitation in the Sierra Nevada range is attributed partly due to vapor fluxes impinging perpendicularly to the Sierras, based on data from ARO wind profilers. However, for individual AR events the HRRR wind

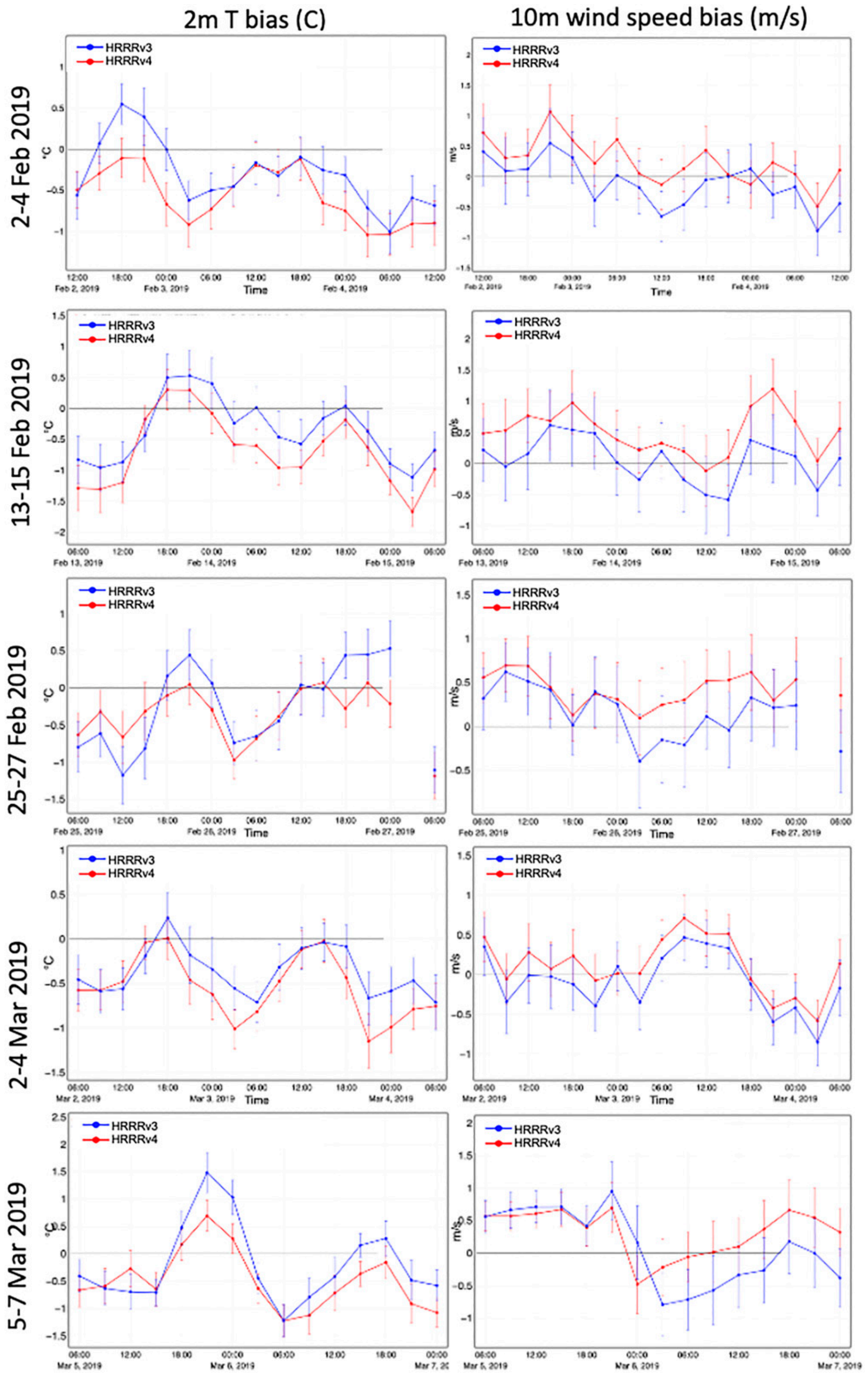

FIG. 10. Time series of HRRRv3 and HRRRv4 (6-h lead times) of 2-m *T* bias and 10-m wind speed bias compared to METAR observations for each AR event (average over AQPI domain).

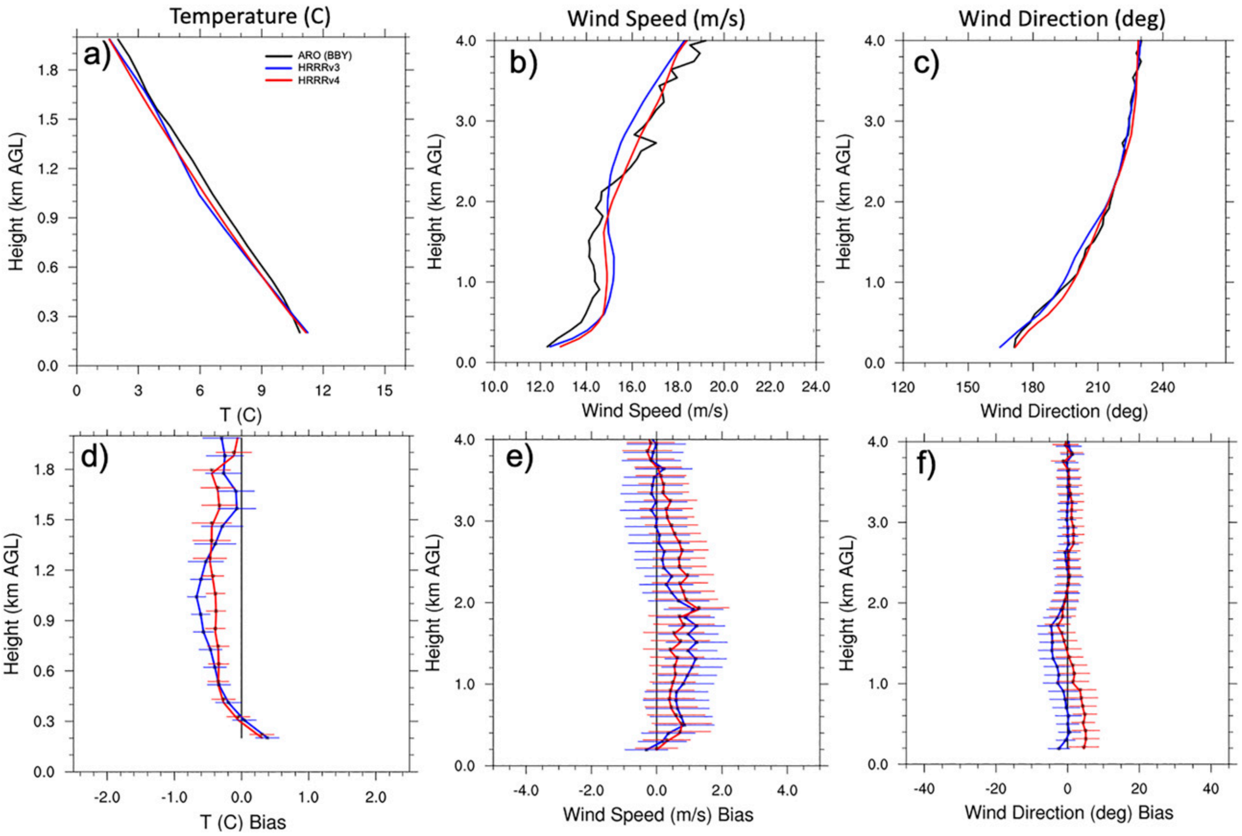

FIG. 11. Profiles of (a)–(c) average and (d)–(f) bias temperature, wind speed, and wind direction at Bodega Bay for the HRRR (6-h lead times) and ARO station, averaged for all five ARs. Error bars represent standard error of the mean at 95% confidence.

speed bias varies (not shown), even though QPF consistently is wetter than QPE products in the Sierra Nevada range, suggesting wind speed errors are not consistently contributing to the QPF/QPE differences. Again, it is difficult to diagnose results from individual events, however, as the differences are not significant at the 95% confidence level.

In the Bay Area, IWV from 6-h forecasts from both HRRRv3 and HRRRv4 are not significantly different than measurements from ARO stations at Bodega Bay and Point Sur (Fig. 13). This suggests that model IWV is not contributing to QPF dry bias in the Bay Area. Darby et al. (2019) also found a reasonable agreement between HRRR and IWV measurements (as well as IWV flux) in the Bay Area. In and near the Sierra Nevada range, both HRRRv3 and HRRRv4 compare well at one location (Old Mammoth), and IWV is consistently smaller than observed at another (Kernville). Therefore, IWV dry biases does not support any model QPF wet biases in the Sierra Nevada range. However, it is possible that errors in the vertical distribution of water vapor could translate to QPF errors, even when IWV compares well.

e. *Comparing HRRR 1–6-h forecast lead times*

Comparing 1-h accumulation forecasts at different lead times can help quantify accuracy of model forecasts at varying lead times and provide some insight into the causes of model QPF errors, especially when comparing the two HRRR model versions—errors at 1-h lead times may indicate errors with model data assimilation or model physics, whereas at longer lead times, the contribution of data assimilation errors wanes and model physics errors may be more likely. Our goals here are to answer two questions: 1) Does the HRRR model generally have better performance at 1- or 6-h lead times? 2) If there are differences between HRRRv3 and HRRRv4, are they more pronounced at 1- or 6-h lead times? We compare to Mesonet only, since Stage IV is only available for 6-h accumulation intervals over the AQPI domain. Furthermore, we compare only when HRRRv4 2-m *T* is between 276 and 400 K, since our Mesonet gauges do not reliably report freezing or frozen precipitation. HRRR biases are larger at 1-h lead times than 6-h over the 0–4200 and 0–1000-m altitude ranges, but smaller at 1-h lead times over the 1000–4200-m altitude range (Fig. 14a). However, CSI is consistently better at 1-h lead times than 6-h over all three altitude ranges (Fig. 14b). These results suggest that different model errors are present at 1- and 6-h lead times, but it is unclear which lead time is more accurate overall. HRRRv4 generally outperforms HRRRv3 at both lead times based on CSI and accumulation bias (Fig. 14), suggesting that model improvements in both data assimilation and physics are likely contributing to forecast improvements. The differences are more pronounced at 1-h lead times, suggesting

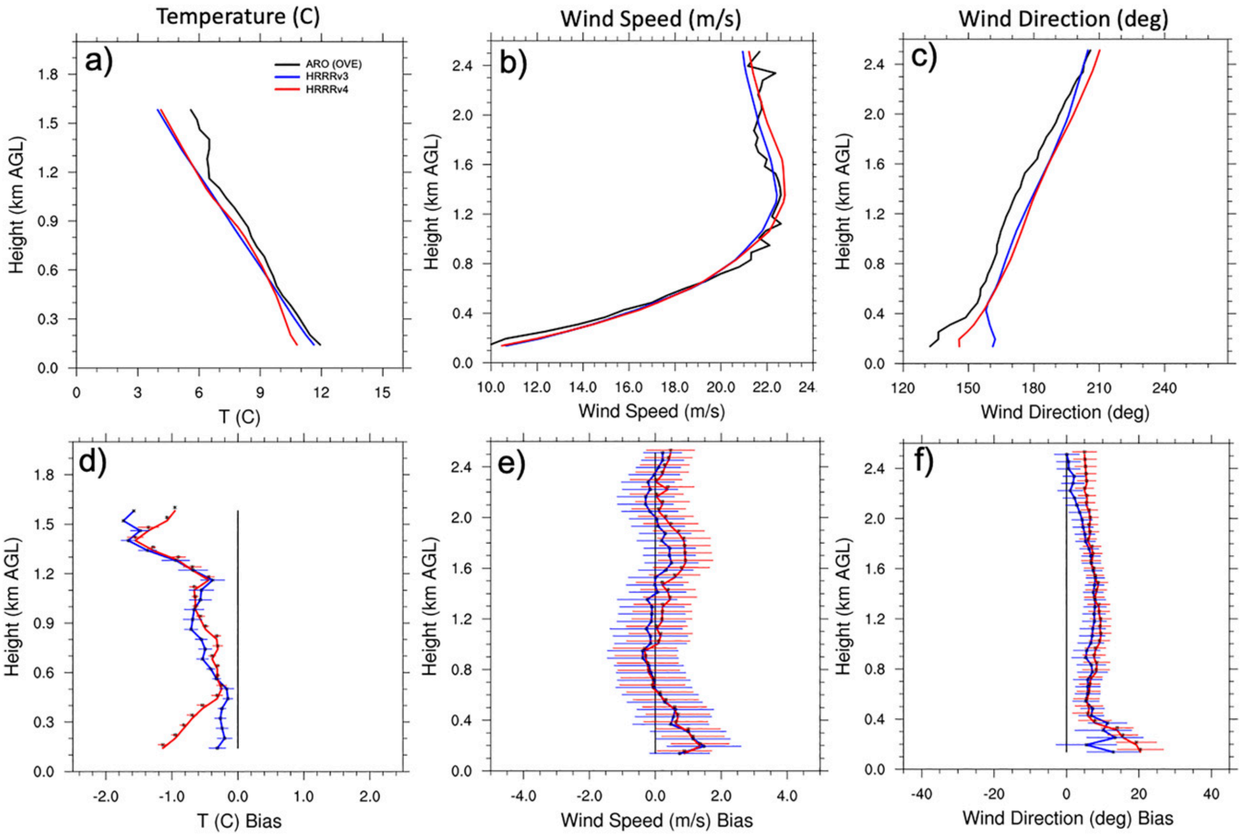

FIG. 12. As in Fig. 11, but at Oroville instead of Bodega Bay.

that HRRRv4 model improvements in data assimilation and/or short time-scale precipitation microphysics may be contributing more to forecast improvements than longer scale physics over HRRRv3, but more detailed experiments are needed to confirm.

f. Evaluating neighborhood maximum precipitation

To investigate whether QPE and QPF differences may be related to small errors in timing/location of precipitation, or whether there are limitations comparing model grids to point gauges, we compare HRRR QPF to Stage IV and Mesonet QPE using the neighborhood max (NM) method in addition to the closest grid box comparisons discussed previously. We conducted preliminary evaluations of the 13–15 February 2019 AR event using neighborhood radii of 3, 10, and 40 km (not shown). Spatial patterns of HRRRv4 QPF biases compared to Mesonet QPE showed similarities regardless of the radius chosen, with HRRRv4 dry biases in the Bay Area and wet biases over the Sierra Range. We conduct a deeper investigation here with a neighborhood radius of 40 km, which is a commonly used radius to study severe weather by the National Weather Service. The HRRRv3 and HRRRv4 CSI at the three thresholds are computed against Stage IV and Mesonet via the 40-km NM method and averaged across all five AR events (Fig. 15). For our analysis at different altitude ranges, the altitude is constrained at each grid box, but QPF from surrounding (neighborhood) grid boxes are not excluded based on the altitude in those grid boxes. Hence, some grid boxes may be assigning QPF from neighboring grid boxes that are outside of the designated altitude ranges. This only occurs in locations where neighboring grid boxes are near the 1000-m altitude cutoff between altitude ranges, such as the slopes of the windward and lee sides of the Sierra Range, and the Coastal Range north of the Bay Area. CSI values for the 40-km NM are generally larger than those computed via closest grid box (Fig. 7), which is to be expected since extending precipitation into a neighborhood allows slight errors in timing/positioning of precipitation to still be considered hits instead of misses. There is less HRRR CSI variation from one event to another via 40-km NM than closest grid box, particularly at higher thresholds, but both methods lead to similar conclusions: 1) HRRRv4 tends to outperform HRRRv3 regardless of the method used; 2) HRRR CSI is larger (better) over the 0–4200-m altitude range than the 1000–4200-m range compared to both Stage IV and Mesonet; and 3) HRRR CSI is larger (better) at lower thresholds than higher thresholds. Performance diagrams show some similarities and some differences between the closest grid box and the 40-km NM method (Fig. 8): As with closest grid box, HRRRv4 tends to outperform HRRRv3, with higher POD, higher success rate, and higher CSI. Additionally, HRRR performance is better versus Stage IV

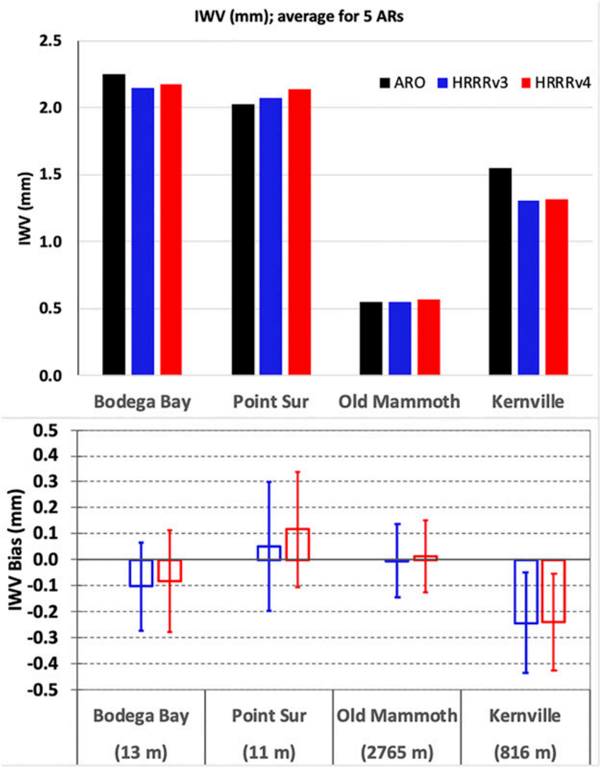

FIG. 13. IWV at ARO stations (black), HRRRv3 (6-h lead times) (blue), and HRRRv4 (6-h lead times) (red) in the Bay Area (Bodega Bay and Point Sur) and the Sierra Nevada area (Old Mammoth and Kernville), averaged across all five AR events. Error bars represent standard error of the mean at 95% confidence.

at low altitude and worse versus Mesonet at high altitude. However, the 40-km NM method confines POD to a narrow region of around 90%–95%, and frequency bias is more consistently large. This makes it more difficult to identify differences between altitude ranges and model versions. Also, there is a bigger difference in CSI between the two model versions when using the 40-km NM method instead of the closest grid box. Therefore, we conclude that closest grid box comparisons are likely adequate when studying HRRR forecast performance of AR events versus Stage IV and Mesonet. However, since this is a single assessment, and the closest grid box and 40-km NM do show some differences, more work should be conducted before concluding that closest grid box comparisons are always adequate for evaluating AR events over the AQPI domain.

4. Discussion and conclusions

We investigated QPE and QPF for five AR events that occurred in the AQPI domain within California/Nevada in February and March 2019. We compared QPF from two versions of the HRRR model (HRRRv3 and HRRRv4) to two QPE products (Stage IV and Mesonet), and compared other HRRR meteorological fields to available measurements. Our findings and recommendations for future work are as follows:

1) The five AR events studied had some spatial similarities in accumulated precipitation: In all five events, accumulated precipitation was generally highest in the Sierra Nevada range, followed by the Bay Area and Pacific Coast, and lowest in the Central Valley. Both QPE products (Stage IV and Mesonet) had general agreement on spatial distribution, although Stage IV had higher precipitation than Mesonet over the Sierra Nevada range, which is partly attributed to the lack of frozen precipitation measured by Mesonet.
2) The spatial distribution of QPF from HRRRv3 and HRRRv4 compared reasonably well to Stage IV and Mesonet, with highest precipitation totals in the Sierra Nevada range, and the lowest precipitation in the Central Valley. However, both HRRRv3 and HRRRv4 were wetter than Stage IV and Mesonet in the Sierra Nevada range for four out of five events, and had a dry bias along the Pacific Coast, particularly in the Bay Area, for all five events.
3) In the Sierra Nevada range, the QPF and QPE products compared well at temperatures above freezing. Below freezing, issues with representation of frozen precipitation in QPE products preclude an accurate assessment of HRRR forecast skill, since our Mesonet database includes liquid precipitation only, and other studies have found Stage IV to poorly estimate frozen precipitation in mountainous terrain. HRRR QPF errors could be significant, and may be due to lower tropospheric wind speed or wind direction biases such as those identified at Oroville. Unfortunately, due to a model cold bias, the HRRR likely produced too much frozen precipitation and too little rain in the Sierra Nevada range, making it difficult to diagnose QPF versus QPE errors by removing frozen precipitation from the HRRR, or by comparing HRRR snow accumulation to snow stations. HRRR IWV mostly compared well with ARO measurements, but was sometimes too low, which contrasts with a general wet bias in this region. More research is needed to distinguish QPF errors from QPE errors in this region, and some ideas are presented in the future work below.
4) In the Bay Area, HRRR IWV, wind speed, and wind direction compared relatively well with ARO measurements. The HRRR temperature profile showed a warm bias near the surface and a cold bias above 500 m, suggesting some errors in model representation of meteorological features such as the low level jet or frontogenesis, which could explain some of the QPF dry bias in the Bay Area. However, temperature and wind speed were quite similar between HRRRv3 and HRRRv4, even though HRRRv4 had reduced precipitation bias and improved CSI/ETS, suggesting other model configuration differences were responsible for the improved forecasts. For a few AR events, HRRRv4 IWV compared better to measurements than HRRRv3, possibly explaining the improved CSI and reduced accumulation bias. Other work finds a consistent QPF dry bias in HRRR and WRF-based models in the Bay Area, and while it is possible that IWV, low level temperature or wind speed biases are contributing to QPF biases, more work is needed to investigate

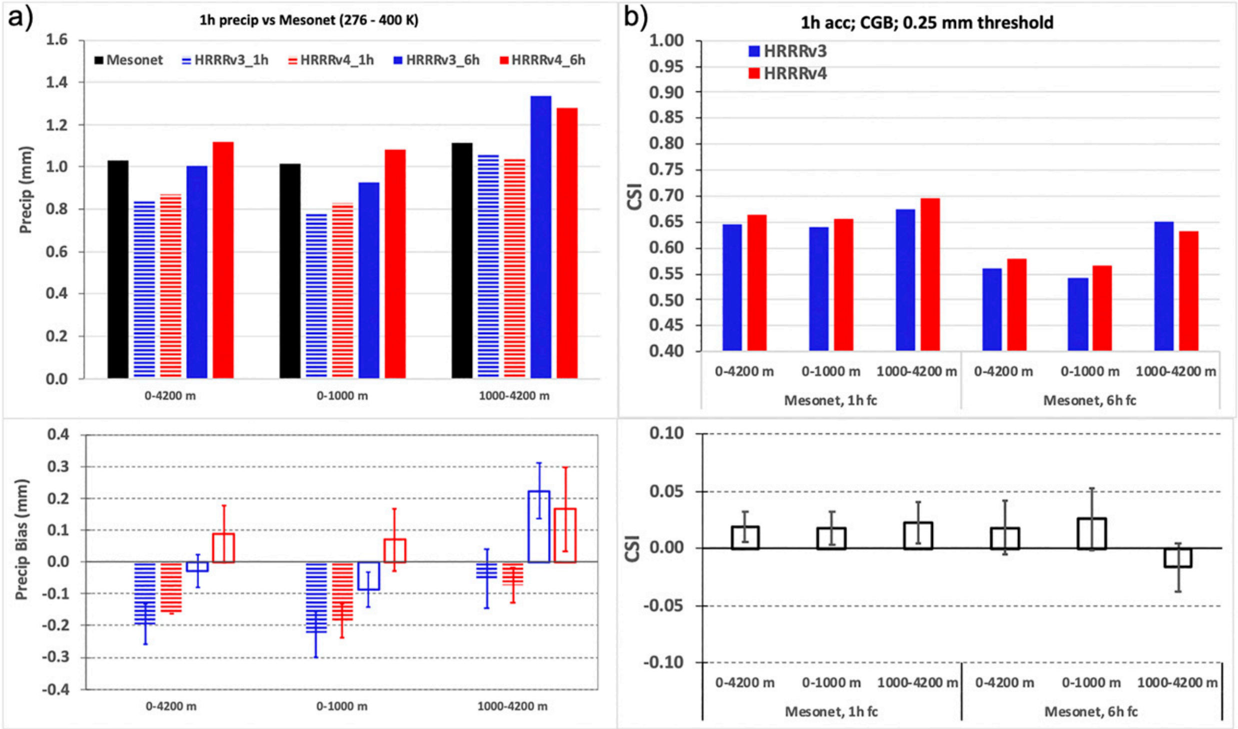

FIG. 14. Average 1-h inferred liquid accumulation (mm) for Mesonet, HRRRv3, and HRRRv4 at two lead times (1 and 6 h), averaged across the AQPI domain for all five AR events, at three different altitude ranges, using only grid boxes where HRRRv4 2-m T is between 276 and 400 K, and data are present in all three datasets. (a) Average 1-h accumulation (mm) and bias. Bar heights in the bottom panel represent precipitation bias relative to Mesonet, and error bars represent 95% confidence interval of the standard error of the mean. (b) CSI at 0.25-mm threshold. Bar heights in the bottom panel are the difference in CSI between HRRRv3 and HRRRv4, with error bars representing the 95% confidence interval for these differences based on the bootstrapping technique (Hamill 1999).

the impacts specific model physics and microphysics parameters have on QPF biases in the Bay Area.

5) HRRRv4 usually outperformed HRRRv3 for the AR events compared to both Stage IV and Mesonet using several precipitation evaluation metrics (accumulated precipitation bias, frequency bias, POD, success rate, CSI, and ETS). QPF was improved at both 1- and 6-h lead times, with larger improvements at 1-h lead times, suggesting that mainly improvements in data assimilation, and also possibly physics, are contributing to improved forecasts

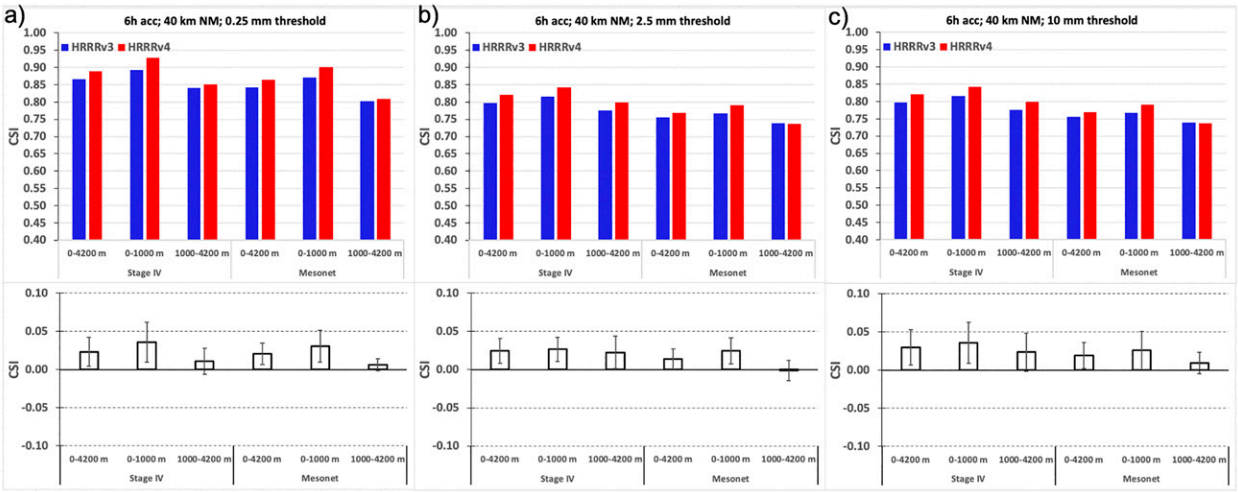

FIG. 15. As in Fig. 7, but using the neighborhood maximum (NM) method (Schwartz 2017) instead of closest grid box.

from HRRRv4 compared to HRRRv3. Wind profile biases were similar between HRRRv3 and HRRRv4, suggesting that low level wind errors may be continuing to cause QPF errors in both model versions, although it is difficult to conclude here as QPF errors may arise from many model parameter deficiencies.

6) While other studies have found advantages of applying neighborhood or object-based methods to precipitation forecast skill for convective storms, we found that evaluating the closest grid box is likely adequate to compare QPE and QPF for AR events. Conclusions were similar when using the 40-km NM method as when using the closest grid box for comparing the HRRRv3 and HRRRv4 to the Stage IV gridded product and to Mesonet point gauges. However, there were some differences between the two methods: The 40-km NM method confines POD to a narrower region, frequency bias is more consistently large, and the difference in CSI is larger between the two model versions. It is possible that the lack of radar usage in Stage IV by the CNRFC reduces the usefulness of the NM method by producing a smoother QPE field. More work should be done with this and other neighborhood techniques to determine their usefulness for AR events over mountainous terrain.

Future work should further explore the causes of the QPF dry bias in the Bay Area reported in this and other research. HRRR representation of water vapor vertical profiles, near-surface winds, and their relationship to meteorological features such as fronts and the low-level jet, as well as deeper investigation of other model physics and microphysics at several forecast lead times could help understand and improve model forecasts of AR events.

Future work should also focus on improving both QPF and QPE of frozen precipitation (which primarily impacts the Sierra Nevada range and the Pacific Coastal mountains). Comparisons of model liquid QPF to rain gauge networks and model frozen QPF to snow networks can provide further insights, particularly when evaluating models with fairly accurate temperature profiles. Evaluations at different temperature thresholds could be repeated with other approaches. More accurate temperature measurements could be used to constrain temperature ranges, such as measurements from the Mesonet network. Temperature profiles could be utilized instead of simply near-surface temperatures to more confidently determine precipitation phase.

Comparisons to additional QPE products such as the MRMS QPE product (Zhang et al. 2011; Wu et al. 2012; Zhang et al. 2016) or the probabilistic QPE product (Bytheway et al. 2021, manuscript submitted to *Wea. Forecasting*) in addition to Stage IV and Mesonet should also provide more useful information. Adding reflectivity from two X-band radars (Cifelli et al. 2018) to the reflectivity mosaic assimilated by the HRRR could also be useful. A deeper look at model performance at varying lead times can help understand contributions of model physics and assimilation errors. Finally, evaluating higher model horizontal and vertical resolution could determine whether limitations in lower atmosphere or terrain representation are contributing to model QPF errors.

Acknowledgments. We thank Eric James and Jaymes Kenyon for help with setting up and troubleshooting the RAP/HRRR experimental retrospective simulations. The ARO data were made available from the NOAA/Physical Sciences Laboratory. Helpful comments from one internal reviewer (Jeff Duda) and three anonymous external reviewers have vastly improved the manuscript. This work is funded in part by the AQPI research program via NOAA Award 3RR2NAQ-P02 and by the NOAA Cooperative Agreement with CIRES, NA17OAR4320101.

Data availability statement. HRRR operational and experimental output and observations used for data assimilation are available on the NOAA High Performance Storage System at their standard folder locations for real-time runs, and at /ESRL/BMC/wrfruc/5year/jenglish/ncep/febmar2019/ for the experimental RAPv5/HRRRv4 retrospective simulations. HRRR operational output is also archived at NCEP as well as Google Cloud at https://console.cloud.google.com/marketplace/product/noaa-public/hrrr. The plotting database used to generate Fig. 10 is available at https://esrl.noaa.gov/gsd/mats/.

REFERENCES

Alexander, C., and Coauthors, 2017: WRF-ARW research to operations update: The Rapid-Refresh (RAP) version 4, High-Resolution Rapid Refresh (HRRR) version 3 and convection-allowing ensemble prediction. *18th WRF User's Workshop*, Boulder, CO, UCAR–NCAR, 2.5, https://ruc.noaa.gov/ruc/ppt_pres/Alexander_WRFworkshop_2017_Final.pdf.

Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, https://doi.org/10.1175/WAF933.1.

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

Bytheway, J. L., M. Hughes, K. Mahoney, and R. Cifelli, 2020: On the uncertainty of high resolution hourly quantitative precipitation estimates in California. *J. Hydrometeor.*, **21**, 865–879, https://doi.org/10.1175/JHM-D-19-0160.1.

Cannon, F., F. M. Ralph, A. M. Wilson, and D. P. Lettenmaier, 2017: GPM satellite radar measurements of precipitation and freezing level in atmospheric rivers: Comparison with ground-based radars and reanalyses. *J. Geophys. Res. Atmos.*, **122**, 12 747–12 764, https://doi.org/10.1002/2017JD027355.

——, J. M. Cordeira, C. W. Hecht, J. R. Norris, A. Michaelis, R. Demirdjian, and F. M. Ralph, 2020: GPM satellite radar observations of precipitation mechanisms in atmospheric rivers. *Mon. Wea. Rev.*, **148**, 1449–1463, https://doi.org/10.1175/MWR-D-19-0278.1.

Cifelli, R., V. Chandrasekar, H. Chen, and L. E. Johnson, 2018: High resolution radar quantitative precipitation estimation in the San Francisco Bay Area: Rainfall monitoring for the urban environment. *J. Meteor. Soc. Japan*, **96A**, 141–155, https://doi.org/10.2151/jmsj.2018-016.

Clark, A. J., W. A. Gallus, and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, https://doi.org/10.1175/2010WAF2222404.1.

Corringham, T. W., F. M. Ralph, A. Gershunov, D. R. Cayan, and C. A. Talbot, 2019: Atmospheric rivers drive flood damages in the western United States. *Sci. Adv.*, **5**, eaax4631, https://doi.org/10.1126/sciadv.aax4631.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and 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.

Darby, L. S., A. B. White, D. J. Gottas, and T. Coleman, 2019: An evaluation of integrated water vapor, wind, and precipitation forecasts using water vapor flux observations in the western United States. *Wea. Forecasting*, **34**, 1867–1888, https://doi.org/10.1175/WAF-D-18-0159.1.

Davis, A. C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

DeFlorio, M. J., D. E. Waliser, B. Guan, D. A. Lavers, F. M. Ralph, and F. Vitart, 2018: Global assessment of atmospheric river prediction skill. *J. Hydrometeor.*, **19**, 409–426, https://doi.org/10.1175/JHM-D-17-0135.1.

Dettinger, M., 2011: Climate change, atmospheric rivers and floods in California—A multimodel analysis of storm frequency and magnitude changes. *J. Amer. Water Resour. Assoc.*, **47**, 514–523, https://doi.org/10.1111/j.1752-1688.2011.00546.x.

——, 2013: Atmospheric rivers as drought busters on the U.S. West Coast. *J. Hydrometeor.*, **14**, 1721–1732, https://doi.org/10.1175/JHM-D-13-02.1.

Donaldson, R. J., R. M. Dyer, and R. M. Kraus, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints, *Ninth Conf. on Severe Local Storms*, Norman, OK, Amer. Meteor. Soc., 321–326.

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

Gershunov, A., and Coauthors, 2019: Precipitation regime change in Western North America: The role of Atmospheric Rivers. *Sci. Rep.*, **9**, 9944, https://doi.org/10.1038/s41598-019-46169-w.

Gilbert, G. K., 1884: Finley's tornado predictions. *Amer. Meteor. J.*, **1**, 166–172.

Gimeno, L., R. Nieto, M. Vázquez, and D. A. Lavers, 2014: Atmospheric rivers: A mini-review. *Front. Earth Sci.*, **2**, 2.1–2.6, https://doi.org/10.3389/feart.2014.00002.

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.

Guan, B., N. P. Molotch, D. E. Waliser, E. J. Fetzer, and P. J. Neiman, 2010: Extreme snowfall events linked to atmospheric rivers and surface air temperature via satellite measurements. *Geophys. Res. Lett.*, **37**, L20401, https://doi.org/10.1029/2010GL044696.

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

Hatchett, B. J., and Coauthors, 2020: Observations of an extreme atmospheric river storm with a diverse sensor network. *Earth Space Sci.*, **7**, e2020EA001129, https://doi.org/10.1029/2020EA001129.

James, E. P., and S. G. Benjamin, 2017: Observation system experiments with the hourly updating Rapid Refresh model using GSI hybrid ensemble–variational data assimilation. *Mon. Wea. Rev.*, **145**, 2897–2918, https://doi.org/10.1175/MWR-D-16-0398.1.

Jeworrek, J., G. West, and R. Stull, 2021: WRF precipitation performance and predictability for systematically varied parameterizations over complex terrain. *Wea. Forecasting*, **36**, 893–913, https://doi.org/10.1175/WAF-D-20-0195.1.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2011: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons, 274 pp., https://doi.org/10.1002/9781119960003.

Kim, D., B. Nelson, and D. Seo, 2009: Characteristics of reprocessed Hydrometeorological Automated Data System (HADS) hourly precipitation data. *Wea. Forecasting*, **24**, 1287–1296, https://doi.org/10.1175/2009WAF2222227.1.

Kingsmill, D. E., P. J. Neiman, F. M. Ralph, and A. B. White, 2006: Synoptic and topographic variability of Northern California precipitation characteristics in landfalling winter storms during CALJET. *Mon. Wea. Rev.*, **134**, 2072–2094, https://doi.org/10.1175/MWR3166.1.

Konrad, C. P., and M. D. Dettinger, 2017: Flood runoff in relation to water vapor transport by atmospheric rivers over the western United States, 1949–2015. *Geophys. Res. Lett.*, **44**, 11 456–11 462, https://doi.org/10.1002/2017GL075399.

Lavers, D. A., D. E. Waliser, F. M. Ralph, and M. D. Dettinger, 2016: Predictability of horizontal water vapor transport relative to precipitation: Enhancing situational awareness for forecasting western U.S. extreme precipitation and flooding. *Geophys. Res. Lett.*, **43**, 2275–2282, https://doi.org/10.1002/2016GL067765.

——, and Coauthors, 2020: Forecast errors and uncertainties in Atmospheric Rivers. *Wea. Forecasting*, **35**, 1447–1458, https://doi.org/10.1175/WAF-D-20-0049.1.

Lin, Y., and K. E. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. Preprints. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2, https://ams.confex.com/ams/pdfpapers/83847.pdf.

Lundquist, J., M. Hughes, E. Gutmann, and S. Kapnick, 2019: Our skill in modeling mountain rain and snow is bypassing the skill of our observational networks. *Bull. Amer. Meteor. Soc.*, **100**, 2473–2490, https://doi.org/10.1175/BAMS-D-19-0001.1.

Mittermaier, M., N. Roberts, and S. A. Thompson, 2013: A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteor. Appl.*, **20**, 176–186, https://doi.org/10.1002/met.296.

Neiman, P. J., M. Hughes, and B. J. Moore, 2013: Sierra barrier jets, atmospheric rivers, and precipitation characteristics in Northern California: A composite perspective based on a network of wind profilers. *Mon. Wea. Rev.*, **141**, 4211–4233, https://doi.org/10.1175/MWR-D-13-00112.1.

Nelson, B. R., O. P. Prat, D. Seo, and E. Habib, 2016: Assessment and implications of NCEP stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31**, 371–394, https://doi.org/10.1175/WAF-D-14-00112.1.

Ralph, F. M., P. J. Neiman, and G. A. Wick, 2004: Satellite and CALJET aircraft observations of atmospheric rivers over the eastern North Pacific Ocean during the winter of 1997/98. *Mon. Wea. Rev.*, **132**, 1721–1745, https://doi.org/10.1175/1520-0493(2004)132<1721:SACAOO>2.0.CO;2.

——, E. Sukovich, D. Reynolds, M. Dettinger, S. Weagle, W. Clark, and P. J. Neiman, 2010: Assessment of extreme quantitative precipitation forecasts and development of regional extreme event thresholds using data from HMT-2006 and COOP observers. *J. Hydrometeor.*, **11**, 1286–1304, https://doi.org/10.1175/2010JHM1232.1.

——, and Coauthors, 2016: CalWater field studies designed to quantify the roles of atmospheric rivers and aerosols in modulating U.S. West Coast precipitation in a changing climate. *Bull. Amer. Meteor. Soc.*, **97**, 1209–1228, https://doi.org/10.1175/BAMS-D-14-00043.1.

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

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

Schaake, J., A. Henkel, and S. Cong, 2004: Application of PRISM climatologies for hydrologic modeling and forecasting in the western U.S. *18th Conf. on Hydrology*, Seattle, WA, Amer. Meteor. Soc., 5.3, https://ams.confex.com/ams/84Annual/techprogram/paper_72159.htm.

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

Schwartz, C. S., 2017: A comparison of methods used to populate neighborhood-based contingency tables for high-resolution forecast verification. *Wea. Forecasting*, **32**, 733–741, https://doi.org/10.1175/WAF-D-16-0187.1.

Smalley, M., T. L'Ecuyer, M. Lebsock, and J. Haynes, 2014: A comparison of precipitation occurrence from the NCEP stage IV QPE product and the *CloudSat* cloud profiling radar. *J. Hydrometeor.*, **15**, 444–458, https://doi.org/10.1175/JHM-D-13-048.1.

Stein, J., and F. Stoop, 2019: Neighborhood-based contingency tables including errors compensation. *Mon. Wea. Rev.*, **147**, 329–344, https://doi.org/10.1175/MWR-D-17-0288.1.

Stone, R. E., C. A. Reynolds, J. D. Doyle, R. H. Langland, N. L. Baker, D. A. Lavers, and F. M. Ralph, 2020: Atmospheric River Reconnaissance observation impact in the Navy global forecast system. *Mon. Wea. Rev.*, **148**, 763–782, https://doi.org/10.1175/MWR-D-19-0101.1.

Turner, D. D., and Coauthors, 2020: A verification approach used in developing the Rapid Refresh and other numerical weather prediction models. *J. Oper. Meteor.*, **8**, 39–53, https://doi.org/10.15191/nwajom.2020.0803.

Wu, W., D. Kitzmiller, and S. Wu, 2012: Evaluation of radar precipitation estimates from the National Mosaic and Multisensor Quantitative Precipitation Estimation System and the WSR-88D Precipitation Processing System over the conterminous United States. *J. Hydrometeor.*, **13**, 1080–1093, https://doi.org/10.1175/JHM-D-11-064.1.

Zhang, J., and Coauthors, 2011: National Mosaic and multi-sensor QPE (NMQ) system: Description, results and future plans. *Bull. Amer. Meteor. Soc.*, **92**, 1321–1338, https://doi.org/10.1175/2011BAMS-D-11-00047.1.

——, and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.

Zhu, Y., and R. Newell, 1998: A proposed algorithm for moisture fluxes from atmospheric rivers. *Mon. Wea. Rev.*, **126**, 725–735, https://doi.org/10.1175/1520-0493(1998)126<0725:APAFMF>2.0.CO;2.