Demonstrating a Probabilistic Quantitative Precipitation Estimate for Evaluating Precipitation Forecasts in Complex Terrain

JANICE L. BYTHEWAY,[a,b] MIMI HUGHES,[b] ROB CIFELLI,[b] KELLY MAHONEY,[b] and JASON M. ENGLISH[a,c]

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

(Manuscript received 12 May 2021, in final form 26 October 2021)

ABSTRACT: Accurate quantitative precipitation estimates (QPEs) at high spatial and temporal resolution are difficult to obtain in regions of complex terrain due to the large spatial heterogeneity of orographically enhanced precipitation, sparsity of gauges, precipitation phase variations, and terrain effects that impact the quality of remotely sensed estimates. The large uncertainty of QPE in these regions also makes the evaluation of high-resolution quantitative precipitation forecasts (QPFs) challenging, as it can be difficult to choose a reference QPE that is reliable at both high and low elevations. In this paper we demonstrate a methodology to combine information from multiple high-resolution hourly QPE products to evaluate QPFs from NOAA's High-Resolution Rapid Refresh (HRRR) model in a region of Northern California. The methodology uses the quantiles of monthly QPE distributions to determine a range of hourly precipitation that correspond to "good," "possible," "underestimated," or "overestimated" QPFs. In this manuscript, we illustrate the use of the methodology to evaluate QPFs for seven atmospheric river events that occurred during the 2016–17 wet season in Northern California. Because the presence of frozen precipitation is often not captured by traditional QPE products, we evaluate QPFs both for all precipitation, and with likely frozen precipitation excluded. The methodology is shown to provide useful information to evaluate model performance while taking into account the uncertainty of available QPE at various temporal and spatial scales. The potential of the technique to evaluate changes between model versions is also shown.

KEYWORDS: Complex terrain; Precipitation; Atmospheric river; Forecast verification/skill

1. Introduction

In recent years quantitative precipitation forecasts (QPFs) have been produced with more frequent update cycles and increased spatial and temporal resolution. In the United States, the hourly updating operational High-Resolution Rapid Refresh (HRRR; Benjamin et al. 2016) model produces hourly precipitation forecasts at 3-km grid spacing out to 18 h, and out to 36 h at 0000, 0600, 1200, and 1800 UTC beginning with version 3 in 2018. The HRRR model is updated somewhat regularly, and the latest version (HRRR version 4) became operational in December 2020. To evaluate the quality of high resolution QPFs, quantitative precipitation estimates (QPEs) with commensurately high spatial and temporal resolution are needed. While it is possible to aggregate high resolution QPFs for comparison to lower resolution QPEs, the full value of high resolution forecasts is realized at their native resolution.

Ground-based radar observations can provide frequent updates to precipitation estimates at high spatial resolution, and in the eastern United States, QPE products comprised of observations from ground-based radars or ground-based radars combined with other sensors are frequently used for both QPF evaluation and the validation of independent QPE products (e.g., Bytheway and Kummerow 2015; Cai and Dumais 2015; Clark et al. 2014; Davis et al. 2006; Ebert and Gallus 2009; Gourley et al. 2010). Although there are many uncertainties in radar-based QPE due to the indirect relationship between the measured radar reflectivity and rain rate, ground clutter, and beam-filling effects (Villarini and Krajewski 2009); the advent of dual-polarization radar has greatly improved the detection and removal of nonprecipitating artifacts resulting in improved rainfall estimates (Ryzhkov et al. 2005). The eastern United States also has a relatively high availability of rain gauges to both supplement and complement the radar-based QPE.

Satellite-based precipitation products are also being produced at high spatiotemporal resolution on a global scale, providing subhourly to hourly QPE on the order of 5–10-km grid spacing. Like radar, these satellite-based products are prone to uncertainties due to the indirect relationship between surface precipitation and satellite-observed radiances, spatial resolution of the satellite footprint, and algorithm assumptions (Bellerby and Sun 2005; Kirstetter et al. 2015; Tian and Peters-Lidard 2010).

In the complex terrain of the western United States, obtaining high-resolution QPE is much more complicated: low-level radar observations are often blocked by mountains, gauges are more sparse due to difficulty of placement and lack of infrastructure, and satellite-based QPE products are known to have reduced quality over complex terrain, snow-covered surfaces, and when detecting frozen precipitation (Bartsotas et al. 2018; Derin et al. 2016; Hirpa et al. 2010; Dinku et al. 2008; Timmermans et al. 2019; Beck et al. 2019; Ebert et al. 2007; Tian and Peters-Lidard 2010; Sun et al. 2018; Dinku et al. 2010). Even in relatively well-instrumented areas, orographic influences on precipitation can result in large

Corresponding author: Janice L. Bytheway, Janice.Bytheway@noaa.gov

DOI: 10.1175/WAF-D-21-0074.1

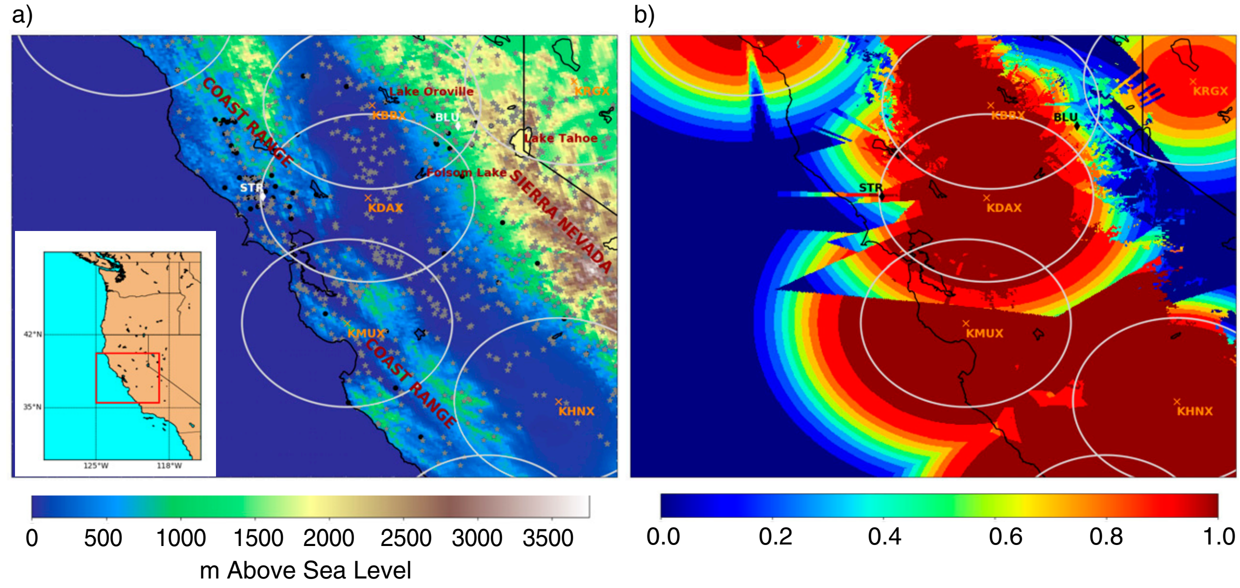

FIG. 1. (a) The AQPI domain, with locations cited in the text labeled. HADS gauges are shown by gray stars, while HMT-West gauges are shown by black dots. The location of WSR-88D radars is given by orange X marks with the station name and 100-km range rings of radars both within and just outside the domain shown with gray circles. Gauge sites used in the text are shown by white diamonds. (b) Example radar quality index from MRMS from 1500 UTC 17 Nov 2017, including radar locations with 100-km range rings and gauge sites cited in the text. Higher values indicate better quality radar estimates.

accumulation differences over relatively small areas that often cannot be resolved by available observation networks. The presence of frozen precipitation further complicates evaluation, as many QPE products underestimate frozen precipitation or do not measure it at all (Lundquist et al. 2019; English et al. 2021b).

The combination of highly heterogeneous precipitation fields and large QPE uncertainty make the evaluation of high-resolution QPFs in regions of complex terrain quite challenging. In fact, some studies have suggested that model simulations may be of higher quality than the observations in these regions (Lundquist et al. 2019). Additionally, large QPE uncertainty can make it difficult to choose a reliable reference product for model validation [i.e., What product (if any) represents the true rainfall in this region?]. Ciach et al. (2007) point out the need for methods that account for errors in QPE datasets that are used to evaluate other QPE products or QPFs. They, as well as Villarini et al. (2009a,b) developed and demonstrated a method that accounts for the uncertainty in radar-derived rainfall estimates when they are used to evaluate QPE or QPF. These methods rely on a dense network of high-quality rain gauges, and would therefore likely be less successful in regions of complex terrain.

One area where complex topography presents challenges to both high resolution precipitation estimation and forecasting is in northern California, United States. Northern California, and the San Francisco Bay Area in particular, is densely populated and prone to both flood and drought (Swain et al. 2018). Therefore, accurate high resolution QPFs are necessary for both water supply and flood mitigation management in the region. A collaboration between water management agencies in the San Francisco Bay Area and the National Oceanic and Atmospheric Administration (NOAA) called the Bay Area Advanced Quantitative Precipitation Information (AQPI) project aims to improve observations and forecasts of precipitation in this region (Cifelli et al. 2018). The AQPI study domain contains several topographic features that challenge both observation and prediction of precipitation, including, from west to east, the transition from sea to land, the Coastal Range rising sharply inland from the coast to elevations of 2000 m, the large Central Valley, and finally the Sierra Nevada mountain range, which reaches elevations of over 4000 m (Fig. 1a). The AQPI domain has served as a test bed for a number of HRRR model experiments that aim to improve QPFs in the region. These include testing a 1-km nest (i.e., higher spatial resolution), testing the impact of an increased number of vertical levels, and data assimilation experiments (English et al. 2021a). With such large QPE uncertainties in this domain (Bytheway et al. 2020), finding a consistently reliable reference with which to evaluate these experimental forecasts at their native resolution has been difficult.

Elsewhere in the United States, particularly east of 105°W, the hourly, 4-km National Centers for Environmental Prediction (NCEP) Stage IV Multisensor QPE (Lin and Mitchell 2005; Nelson et al. 2016) or 1-km Multi-Radar Multi-Sensor (MRMS; Zhang et al. 2011, 2014, 2016) QPE products are often used for QPF validation. Due to the lack of a sufficient gauge network and frequently blocked radar observations (Maddox et al. 2002), the California/Nevada River Forecast Center (CNRFC) did not produce an hourly Stage IV in the

AQPI domain during water year 2017, which is the focus of this study. While hourly MRMS is available in this region during the period of study, it too suffers from uncertainties in northern California due to poor observational data coverage (Bytheway et al. 2019).

In this paper, we propose a QPF evaluation methodology that takes into account the large QPE uncertainty in regions of complex terrain, using the AQPI domain (Fig. 1) as a testbed. The methodology incorporates numerous available high resolution QPE products, acknowledging that all have strengths and weaknesses in regions of complex terrain. The developed methodology relies on the assumption that the "true" QPE is likely within the range of the various estimates, applying qualitative descriptors of forecast performance that depict how the QPF fits within the uncertainty of the QPE rather than calculating traditional evaluation metrics that compare QPF to an absolute reference. The authors acknowledge that this is a big assumption, particularly in high terrain where frozen precipitation dominates. For this reason, in section 4 we perform our forecast evaluation both with and without likely frozen precipitation included. In sections 2 and 3, we describe the available QPE and QPF products and the QPF validation methodology. In section 4 we demonstrate how the methodology is used to evaluate both a single version of the HRRR model and compare changes between two versions of the model. Several case studies are assessed to illustrate common performance characteristics that can be used to inform future model development. Discussion and conclusions are presented in sections 5 and 6, respectively.

2. Data

a. *Quantitative precipitation estimates*

Bytheway et al. (2020) examined the large degree of uncertainty in high-resolution (hourly, <10-km grid spacing) QPE in northern California, identifying ten available products for evaluation that are used in this study as well. This includes four satellite products: the NOAA Climate Prediction Center (CPC) morphing technique (CMORPH; Joyce et al. 2004; Xie et al. 2017) version 1.0, Integrated Multisatellite Retrievals for Global Precipitation Measurement (GPM) (IMERG) version 6 research/final run (Huffman et al. 2018; Tan et al. 2019), Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks–Cloud Classification System (PERSIANN-CCS; Hsu et al. 1997; Hong et al. 2004), and Global Satellite Mapping of Precipitation (GSMaP; Kubota et al. 2007; Ushio et al. 2009). Also included are two precipitation gauge datasets—the Hydrometeorological Automated Data System (HADS; Kim et al. 2009) and gauges operated in the region as part of the Hydrometeorology Testbed–West (HMT, https://hmt.noaa.gov/), and two gauge-informed products—the MRMS Mountain Mapper (MRMS-MM) and MRMS Gauge-Adjusted Radar (MRMS-GA) (Zhang et al. 2011, 2014, 2016). Finally, two multisensor products produced by NCEP are also included—the Real Time Mesoscale Analysis and the Unrestricted Mesoscale Analysis [RTMA (De Pondeca et al. 2011) and URMA, respectively]. Brief descriptions of the available high-resolution QPE follow.

1) SATELLITE-BASED PRODUCTS

(i) CMORPH version 1.0

CMORPH combines instantaneous precipitation estimates retrieved from passive microwave remote sensing on low Earth orbiting (LEO) satellites with motion vectors derived from geostationary infrared (GEO-IR) satellite imagery. As each LEO retrieval is performed, it is interpolated both forwards and backward in time with the derived motion vectors (i.e., the "morphing" process; Joyce et al. 2004), determining the shape and location of precipitating features during periods of time between LEO estimates. Version 1.0 includes a bias correction using daily gauge analysis from the CPC (Xie et al. 2017) and is available globally at 30-min time steps and 8-km grid spacing. Hourly precipitation estimates are obtained by assuming constant rain rates for the duration of each time step and accumulating two half-hourly estimates to a single hour. CMORPH Version 1.0 can be obtained from https://www.cpc.ncep.noaa.gov/products/janowiak/cmorph_description.html.

(ii) PERSIANN-CCS

The PERSIANN-CCS algorithm uses a cloud classification system to categorize cloud features observed by GEO-IR satellites based on height, areal extent, and texture. These categories are then used to assign a rain rate to each pixel within the cloud feature. The rain rates are assigned via an empirical relationship between brightness temperature and rain rate that is both regionally dependent and temporally evolving. Both the cloud classification procedure and precipitation distribution for each cloud type were developed and trained for an artificial neural network using observations from the Tropical Rainfall Measurement Mission (TRMM) satellite and ground-based radar (Hsu et al. 1997; Hong et al. 2004). PERSIANN-CCS estimates are available hourly at 0.04° (approximately 4 km) grid spacing from https://chrsdata.eng.uci.edu/.

(iii) IMERG V06 final run

IMERG is similar to CMORPH, in that precipitation estimates from passive microwave satellites are interpolated between individual LEO satellite overpasses. The IMERG algorithm includes an intercalibration between all of the observed passive microwave radiances prior to performing precipitation retrievals to account for differences in scan strategy, available channels, and overpass times. The microwave-based precipitation estimates are then interpolated between overpasses using the CMORPH technique as well as PERSIANN. Finally, monthly gauge data from a variety of sources is applied to reduce biases (Huffman et al. 2018). IMERG precipitation estimates are available every 30 min at 0.1° (approximately 10 km) grid spacing and can be obtained from https://pmm.nasa.gov/data-access/downloads/gpm.

(iv) GSMAP

GSMAP is produced by the Japanese Aerospace Agency (JAXA) and, similar to CMORPH, combines multiple passive microwave-based precipitation estimates with motion vectors from GEO-IR to produce hourly precipitation estimates at 0.1° (~10 km) grid spacing. In addition to motion vectors, a

Kalman filter technique is also employed that uses cloud-top height measurements from the GEO-IR to estimate changes in precipitation intensity, location, and shape between the LEO overpasses. GSMaP data can be found via https://sharaku.eorc.jaxa.jp/GSMaP_NOW/index.htm.

2) Gauge products

(i) HADS

The HADS gauge dataset includes approximately 7000 gauges operated by multiple state and federal agencies (Kim et al. 2009). The data are monitored as they are ingested for obvious issues such as instrument malfunction or transmission errors and verified to be valid at the top of the hour. If missing values can be proven to have occurred when no rain was present, they are replaced with zeros. Gauges included in the HADS dataset may be used for bias correction or the creation of other QPE products included in this study (i.e., the QPE are not entirely independent). HADS data are archived at the National Centers for Environmental Information (NCEI), and the location of 231 HADS gauges available within the AQPI domain are shown by gray stars in Fig. 1a.

(ii) HMT-West

NOAA's Physical Sciences Laboratory (PSL) operates a number of precipitation gauges in the western United States as part of the NOAA Hydrometeorology Testbed (HMT; https://hmt.noaa.gov/). Though far fewer in number than the HADS gauges, many of the 49 HMT network gauges in the AQPI domain were sited specifically to monitor precipitation interactions with the complex terrain, including 25 gauges in the Russian River basin and 8 in the American River basin above Folsom Lake. The HMT gauges are not included in the HADS network, and therefore represent a completely independent source of QPE, having no influence on any other QPE product. The gauges are maintained by PSL staff and are indicated by black dots in Fig. 1a. Data from these gauges can be obtained from https://psl.noaa.gov/data/obs/datadisplay/.

3) Gauge-informed products (MRMS)

MRMS products are produced at NCEP and combine precipitation estimates from the HADS gauge network, U.S. and Canadian operational radar networks, and the Precipitation-elevation Regression on Independent Slopes Model (PRISM) climatology (Daly et al. 1994, 2017) to produce hourly QPE over the continental United States (CONUS) on a 1-km grid (Zhang et al. 2011, 2014, 2016). Four MRMS products are available: Radar Only, Gauge Only, Gauge-Adjusted Radar, and Mountain Mapper. We focus here on the evaluation of the products that include multiple types of precipitation information: Gauge-Adjusted Radar and Mountain Mapper, which have been shown to perform more favorably in the region than the Radar Only and Gauge Only products (Willie et al. 2017). An updated version of the MRMS product (Version 12) that became operational in October 2020 has yet to be evaluated with respect to the other QPE products in this domain. MRMS data are available from https://www.nssl.noaa.gov/projects/mrms/.

(i) MRMS-Gauge-Adjusted radar

MRMS-GA uses reflectivity from the U.S. WSR-88D network and C-band radars from Environment Canada, identifying different precipitation types that are assigned one of five different reflectivity-rainfall relationships (Zhang et al. 2016). As discussed, radar in the western United States is not always reliable due to beam blockage and inadequate coverage. Radars in the AQPI domain are shown by orange X marks in Fig. 1a, with 100-km range rings of radars within and just outside the domain also highlighted. A radar quality index (Fig. 1b) is employed in the MRMS-GA algorithm to determine where the radar observations may not be reliable, and dual-polarization variables are used to filter out nonprecipitation echoes. Radar-based precipitation estimates are adjusted using the HADS gauge network as described in Zhang et al. (2011).

(ii) MRMS Mountain Mapper

Unlike MRMS-GA, MRMS-MM does not include radar, relying solely on HADS rain gauge data, interpolated via inverse distance weighting to the 1-km grid. MRMS-MM uses the PRISM climatology (Daly et al. 1994, 2017) which employs an elevation model to calculate the linear relationship between precipitation and elevation at monthly and annual scales. These relationships are used to adjust the interpolated gauge-based precipitation estimates for orographic enhancement to produce the MRMS-MM product.

4) Multisensor analyses

The National Weather Service established a program to create a "Reanalysis of Record" in 2004, which is intended to provide analyses at high spatial and temporal resolution over a 30-yr period. The first phase of this program produces the RTMA product, a two-dimensional variational (2DVAR) analysis of 2-m temperature, specific humidity, dewpoint, and 10-m U and V wind components with the 13-km Rapid Update Cycle (RUC) model (or the 13-km Rapid Refresh (RAP) model after 2012) serving as a first guess (De Pondeca et al. 2011). RTMA precipitation estimates are the "early" version of the NCEP Stage II multisensor precipitation analysis bilinearly interpolated to a 2.5-km grid. This analysis is comprised of data from 150 operational WSR-88D radars and gauge data from approximately 1500 Automated Surface Observing System (ASOS) stations and is produced at 35 min past the hour. As not all radar and gauge data are available with such low latency, the analysis is reproduced later with additional data, resulting in the URMA product.

b. High-Resolution Rapid Refresh model

The HRRR model is a convection-allowing model producing hourly forecasts over the continental United States (CONUS) at 3-km grid spacing with hourly updates. HRRR forecasts are produced at 50 vertical pressure levels using the 13-km RAP mesoscale model for boundary conditions (Benjamin et al. 2016). The HRRR model has a similar

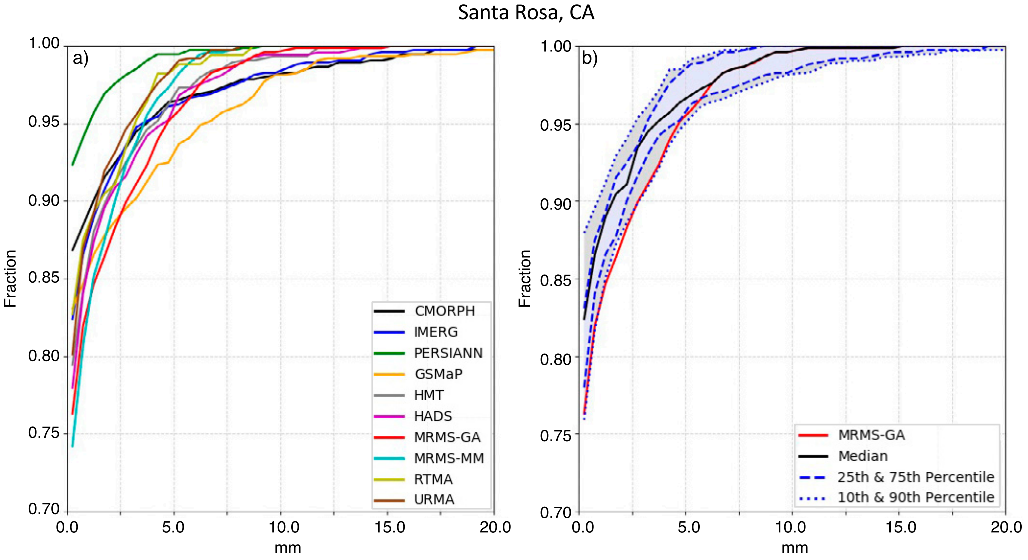

FIG. 2. (a) CDF of 10 QPE products for the grid box closest to Santa Rosa, CA (STR in Fig. 1a), for January 2017. (b) Median, 10th, 25th, 75th, and 90th percentile CDFs of the QPE products shown in (a). IQR is shaded lavender, while the area between the 10th and 90th percentiles is shaded gray. Also shown in (b) is the CDF of the MRMS-GA product for reference.

configuration to the RAP, which is described in Benjamin et al. (2016). The HRRR is continuously under development at the NOAA/ESRL Global Systems Laboratory (GSL), where an experimental version of the model is run in real time alongside the operational version and archived for comparison. New versions of the model become operational approximately every two years.

3. Design and methodology

a. *Development of probabilistic QPE*

As in Bytheway et al. (2020), the gridded QPE datasets described in section 2 are linearly interpolated to the 3-km grid spacing of the HRRR model in order to allow for direct comparison with the QPFs. Gauge data are matched to HRRR model grid boxes via nearest-neighbor, and when multiple gauges are present in a 3-km grid box, their average is taken to represent that location. The results of the Bytheway et al. (2020) study showed that each of the ten QPE products had both strengths and weaknesses when estimating hourly precipitation, but that overall, the uncertainty in hourly QPE is quite large in the AQPI domain. Here, we develop a method to evaluate QPFs from the HRRR model using a probabilistic QPE (pQPE) to classify forecast performance into one of four categories.

We begin by developing a cumulative distribution (CDF) of hourly rainfall in each 3-km grid box from each individual QPE product for each month (i.e., a CDF for January, one for February, etc.). Monthly distributions are chosen for several reasons. First, they are representative of the observed precipitation that occurred in that grid box over the course of the month—if it is a period of particularly heavy rain, those values will be captured in the monthly distribution (likewise, the CDF for a month dominated by relatively light precipitation will not include heavy rain rates). Second, it allows for an adequate number of precipitating hours to be used in the creation of the distribution. CDFs on shorter (e.g., weekly or event based) time scales were also tested, and while they better represented individual storm characteristics, they produced much noisier results and are not as useful during dry periods or during shorter events. Monthly CDFs also provide a statistical constraint on the precipitation that is useful in the event of a timing mismatch between products, or if a gauge that is used to inform or bias-adjust other products malfunctions and produces a large over or under estimate. The monthly CDF of rainfall would dampen this signal, reducing the potential for large outliers.

The CDFs of all QPE products available in the closest 3-km grid box to Santa Rosa, California (STR, Fig. 1a), are shown in Fig. 2a for January 2017. From these CDFs, the median, interquartile range (IQR, 25th percentile and 75th percentile), and 10th and 90th percentile CDFs are determined for each rain rate, as shown in Fig. 2b. These CDFs are calculated for each HRRR 3-km grid box and month of the study period.

Next, a reference QPE for the evaluation is chosen. A reference is necessary because the CDFs shown in Fig. 2 are calculated monthly, while we are interested in evaluating hourly precipitation forecasts. Therefore, we need one of the QPE products to provide an initial guess as to whether is it precipitating at the given location and time, and if so, how much. This initial guess identifies the quantile of the median all-QPE CDF that will be used to determine the range of QPE values that define forecast performance categories. For this demonstration, we have chosen the MRMS-GA product

TABLE 1. Statistical comparison between QPE products and HADS (HMT-West) gauges for all of 2017. Smallest RMSE and bias values and highest correlation coefficient are highlighted in bold.

	RMSE (mm)	Bias	Correlation
CMORPH	0.99 (1.04)	−0.13 (−0.12)	0.35 (0.43)
IMERG	0.73 (0.77)	−0.20 (−0.19)	0.48 (0.55)
GSMaP	1.11 (1.13)	−0.21 (+0.19)	0.39 (0.47)
PERSIANN-CCS	0.91 (0.96)	−034 (−0.48)	0.20 (0.21)
MRMS-GA	0.47 (**0.51**)	**−0.01** (−0.06)	0.80 (**0.81**)
MRMS-MM	0.60 (0.54)	+0.02 (**−0.01**)	0.70 (0.79)
RTMA	0.62 (0.58)	−0.46 (−0.47)	0.54 (0.73)
URMA	**0.40** (0.87)	−0.14 (−0.57)	**0.86** (0.49)

for our reference because it produces the best statistical comparison with both the HADS and HMT gauges (Table 1). The agreement between MRMS-GA and HADS is somewhat expected, since HADS gauges are used in producing the MRMS-GA product. However, there are known conditions when the MRMS-GA is unreliable, including areas where radar may be blocked and HADS gauges are sparse (Bytheway et al. 2019), where frozen precipitation may be occurring (English et al. 2021b) or at high elevations. For example, at elevations less than 1000 m, the MRMS-GA compares more favorably to HADS (RMSE = 0.38 mm, bias = +0.04, correlation = 0.85) than at elevations exceeding 2000 m (RMSE = 0.73 mm, bias = −0.13, and correlation = 0.59). A similar degradation in performance is also found when comparing MRMS-GA to the HMT gauges at high versus low elevations. While the MRMS-GA is used to provide an initial guess at the hourly precipitation at a given location, this methodology incorporates the CDF of all available QPE products to provide a range of probable hourly precipitation amounts. This is particularly useful in situations such as the one shown in Fig. 2b: the MRMS-GA product produces among the lowest frequency of occurrence of hourly accumulations less than 5.0 mm, while the median of all QPE products indicates these accumulations occur up to 5% more often. STR is located in the Russian River valley, in an area where the lowest elevation scans of the Sacramento, California (KDAX), radar are blocked (Fig. 1). Bytheway et al. (2019) showed that this results in the radar frequently not observing shallow precipitating clouds, producing underestimates by MRMS-GA. By incorporating the information from all of the available QPE into the forecast validation, we can account for such situations when the chosen reference may be less reliable.

b. *Forecast evaluation with pQPE*

Once the reference QPE product is chosen, the forecast evaluation continues as follows and is schematized in Fig. 3, which shows the precipitation quantiles from Fig. 2b. As an example, we demonstrate a case where MRMS-GA indicates an hourly accumulation of 5.0 mm at STR for the hour being evaluated.

Starting with the MRMS-GA reference precipitation estimate for the hour and location of interest (5.0 mm), the percentile of the monthly median CDF corresponding to the MRMS-GA-estimated hourly rainfall is determined. Following the black dashed arrow in Fig. 3, an MRMS-GA hourly rainfall of 5.0 mm corresponds to the ~96th percentile of the monthly median all-QPE CDF. The IQR of the all-QPE CDFs (lavender shading between dashed blue lines) at the 96th percentile corresponds to hourly rain accumulations between ~3.9 and 5.8 mm (dashed green arrows). If the QPF falls within this range of hourly accumulations, it is deemed a "good" forecast that falls within the uncertainty of the QPE (illustrated on Fig. 3 by a green star). The 10th and 90th percentile all-QPE CDFs (grayshading between dotted blue lines) at the 96th percentile correspond to hourly rain accumulations between ~3.5 and ~6.8 mm (solid yellow arrows). These values determine the range of "possible" rainfall for that hour and location. A QPF outside the range considered to be a "good" forecast, but within the range indicated by the 10th and 90th percentile accumulation values, is considered to be a "possible" forecast (illustrated by a yellow star on Fig. 3), i.e., the predicted rainfall is not outside of the realm of possibility given the available QPE, but is closer to the tails of the distribution. QPFs below the "possible" range are deemed underestimates (illustrated by a blue star on Fig. 3), while QPFs above the "possible" range are considered overestimates (illustrated on Fig. 3 by a red star). The "stoplight" color scheme (green, yellow, red, and blue) will be used to denote "good," "possible," "overestimated," and "underestimated" forecasts, respectively, for the remainder of this manuscript.

c. *Atmospheric river cases*

HRRR QPFs for seven atmospheric river events that impacted the AQPI domain during the 2016–17 wet season (October–March) are evaluated. The HRRR Version 2 (HRRRv2) was the operational version during this period, while HRRR Version 3 was also being produced in experimental mode (HRRRx). The cases selected for comparison here use a version of the HRRRx that was still under development and therefore may have included slightly different configurations for each case. Since the goal of this manuscript is to demonstrate the utility of the developed methodology, sensitivity of the results to these configuration changes are not examined. Cases were selected in order to compare the performance of the HRRRv2 with the HRRRx. Over the duration of each case, we compare QPFs from forecasts initialized every 3, 6, and 12 h starting at 0000 UTC (e.g., an event lasting 24 h consists of eight consecutive 3-h forecasts, four consecutive 6-h forecasts, and two consecutive 12-h forecasts). In this way we can examine both forecast performance and the changes between model versions at different lead times. Case selection was somewhat limited because HRRRx was not produced at every possible initialization hour, and missing data due to data feed issues or system maintenance are not backfilled. Therefore, some hours of the atmospheric river events are not included in the analysis (i.e., the evaluation begins several hours after (before) the onset (cessation) of the precipitation). Table 2 lists the cases considered.

To demonstrate the various ways the pQPE evaluation methodology can be used to communicate QPF performance to both

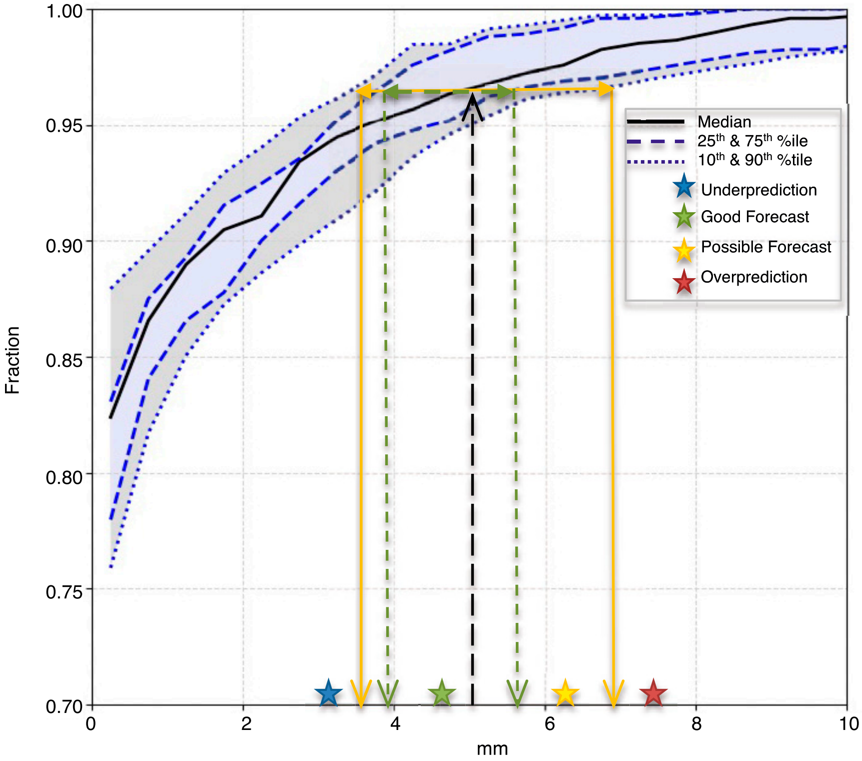

FIG. 3. As in Fig. 2b, illustrating how the model evaluation methodology is performed. Dashed green and solid yellow arrows show the range of hourly precipitation defined as "good" (between the 25th and 75th percentiles) and "possible" (between the 10th and 90th percentiles), respectively.

forecast users and model developers, we first present the results of an evaluation of both versions of the HRRR for a single atmospheric river event that occurred from 1200 UTC 18 February to 1000 UTC 21 February 2017. Aggregate results from all cases will also be discussed. Finally, we examine the effects of removing frozen precipitation from consideration.

4. Results

a. *Overall evaluation*

Figure 4 shows the event total precipitation from both HRRRv2 and HRRRx for 1–3-, 1–6-, and 1–12-h forecasts (i.e., a new forecast initialized every 3, 6, or 12 h), as well as the relative difference between the two. In the 3-h forecasts, the difference between versions of the HRRR are relatively small. South of the Bay Area, HRRRx predicts less precipitation over the course of this event at all forecast lengths. Both versions of the HRRR produce more precipitation in the longer lead forecasts, with HRRRx displaying a larger increase in domain mean rainfall at longer forecast lengths. In several locations, HRRRx 12-h forecasts increase event total precipitation by over 150% compared to HRRRv2. Many of these large changes appear to be related to either increases in very small amounts of rainfall, as in the northernmost part of the domain, or small shifts in the location of bands of higher accumulations, such as the east–west band of enhanced precipitation near the San Francisco Bay. While the HRRRx does increase the amount of precipitation in this feature over HRRRv2, it also shifts it slightly to the south to an area where HRRRv2 produced little accumulation.

Figure 5 shows forecast hourly precipitation from both versions of the HRRR, observed hourly precipitation from the MRMS-GA and the QPE uncertainty bounds at Blue Canyon (BLU, Fig. 1a) for the 18–21 February 2017 event. To avoid overpenalizing the model for small location or timing errors in either the edges or start/end of precipitation, when the reference QPE indicates zero precipitation any forecast precipitation

TABLE 2. List of cases evaluated in this study. Results are shown from the 18–21 Feb case in Figs. 4–8, while aggregate results from all cases are shown in Figs. 9–11.

Begin date	End date	Duration (h)
0000 UTC 23 Dec 2016	0000 UTC 25 Dec 2016	48
1700 UTC 8 Jan 2017	0300 UTC 11 Jan 2017	60
0100 UTC 12 Jan 2017	0700 UTC 13 Jan 2017	30
0000 UTC 3 Feb 2017	0000 UTC 11 Feb 2017	192
1200 UTC 18 Feb 2017	1000 UTC 21 Feb 2017	70
0000 UTC 27 Feb 2017	0000 UTC 28 Feb 2017	24
0000 UTC 21 Mar 2017	0000 UTC 23 Mar 2017	48

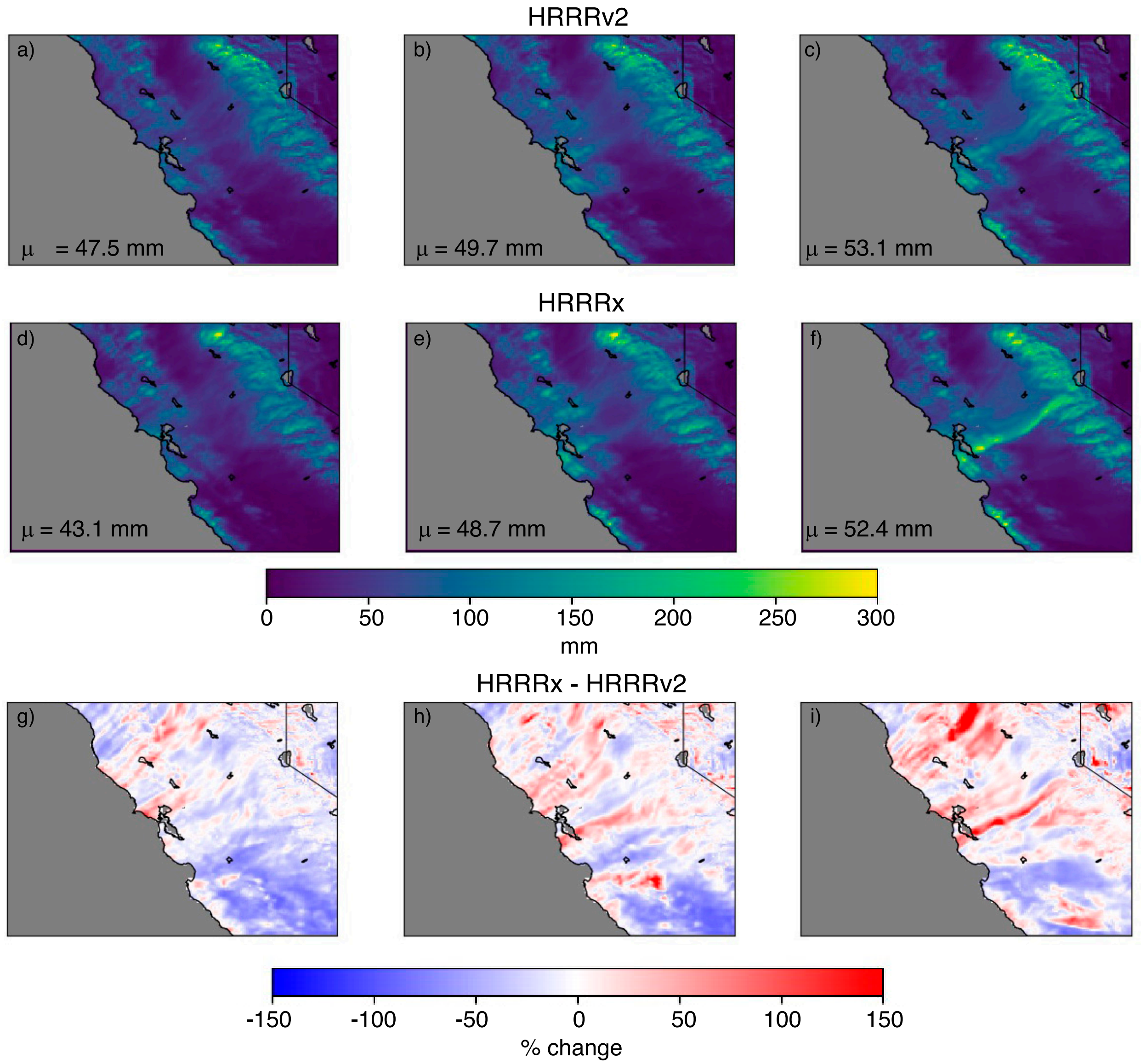

FIG. 4. Event total rainfall from (a),(b),(c) HRRRv2; (d),(e),(f) HRRRx; and (g),(h),(i) relative difference between the two for the atmospheric river event lasting from 1200 UTC 18 Feb to 1000 UTC 21 Feb 2017. Event totals and percent change are calculated using forecasts initialized (left) every 3 h; (center) every 6 h; and (right) every 12 h. Domain mean total rainfall values are shown in (a)–(f) for reference.

less than 0.5 mm is considered a "good" forecast, and any forecast precipitation less than 1 mm h^{-1} is considered "possible."

At BLU, this event can be characterized by two periods of precipitation: from 1200 UTC 18 February to 0400 UTC 19 February, and from 2200 UTC 19 February to 1000 UTC 21 February. Early in the event (~1400 UTC 18 February), the HRRRv2 6- and 12-h forecasts begin producing moderate precipitation too early, resulting in overestimates, while the HRRRx forecasts are in better agreement within the uncertainty of the QPE at this location (Fig. 5 center and bottom). Both versions of the HRRR reduce the hourly precipitation by 0100 UTC 19 February, ending the first period of precipitation several hours too early. During a lull in the second period of precipitation at around 0400 UTC 20 February, the 6- and 12-h forecasts from the HRRRx predict an increase in precipitation several hours prior to it being observed by MRMS-GA. Evidence of this timing error affecting a large part of the domain is apparent in Fig. 6b.

Figure 6a shows the fraction of grid boxes in the domain in each performance category for the HRRRv2. As precipitation intensity increases during the second period of precipitation (~0600 UTC 19 February), the fraction of pixels classified as "good" forecasts decreases sharply, from approximately 80% to nearly 40%, while "possible," "overpredicted," and

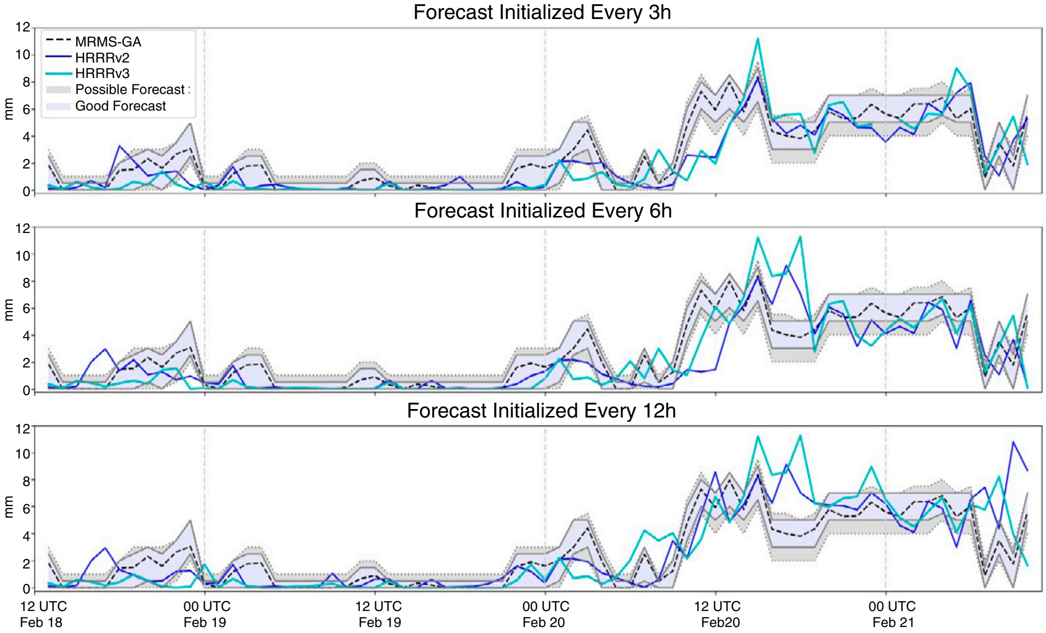

FIG. 5. Time series of hourly rainfall at Blue Canyon (BLU) for the MRMS-GA and (top) 3-, (middle) 6-, and (bottom) 12-h forecast periods from both HRRRv2 and HRRRx for the 18–21 Feb atmospheric river event. Lavender shading indicates the range of "good" QPF values, while gray shading indicates the range of "possible" QPF values. Dashed vertical lines indicate 0000 UTC each day for reference.

"underpredicted" precipitation forecasts each account for approximately 20% of grid boxes in the domain during this period. Figure 6b shows the difference between the HRRRx and HRRRv2, and the aforementioned timing error appears as a large decrease in the amount of grid boxes with hourly QPE characterized as "good" (green lines), with a corresponding large increase in the fraction of grid boxes characterized as "overestimates" (red lines) several hours before the corresponding reduction in performance seen in the HRRRv2 in Fig. 6a.

Figure 6a also shows a reduced fraction of good forecasts with a corresponding rise in overestimates in the first several hours of the event for HRRRv2. During the same period in Fig. 6b, the HRRRx tends to reduce the number of overestimates at all forecast lead times, leading to an increase in the fraction of grid boxes where the forecast precipitation is categorized as "good." After 1000 UTC 20 February, there is a tendency for the HRRRx to produce a larger number of underestimates than the HRRRv2. With the exception of the apparent timing error, the difference between the two model versions in the fraction of grid boxes in each performance category is less than 10% for this event.

The Blue Canyon site is one of a few HRRR grid boxes that have both a HADS and HMT gauge; however, many grid boxes contain no gauges. Figure 7 shows the same results as Fig. 5 with the gauges removed from the QPE uncertainty calculations. In some cases, lack of gauge information reduces the range of the IQR, for example between 1800 UTC 18 February and 0000 UTC 19 February, there are hours where the 3-h HRRRx and 12-h HRRRv2 forecasts are categorized as "good" forecasts in Fig. 5 (top and bottom, respectively), but without the gauge information, the range of both the IQR and the 10th and 90th percentiles decreases, so that these two forecasts are now only rated "possible." A similar reduction in the range of values that constitutes a "good" forecast occurs a few hours before 0000 UTC 20 February. Conversely, at 1200 UTC 20 February, lack of gauge information increased the range of the 10th and 90th percentiles, such that heavy precipitation predicted by the HRRRx went from being categorized as an overestimate to being categorized as "possible."

Figure 8 shows the fraction of time during the event that the HRRRv2 3-h forecast fell into each performance category as well as the difference between the experimental and operational versions. The HRRRv2 predicts more precipitation than the range of QPE in the high terrain of the Sierra Nevada up to 50% of the time (Fig. 8f) and is very rarely categorized as an underestimate in this area (Fig. 8e). On the other hand, the HRRRv2 QPF underestimates compared to all available QPE 20%–30% of the time over most of the rest of the domain. This low bias is consistent with other studies evaluating the HRRR and other WRF-based models (Darby et al. 2019; Dougherty et al. 2021; English et al. 2021b). The HRRRx increases the number of hours of "good" forecasts in the Sierra Nevada (Fig. 8c), with a corresponding reduction of overestimates in this region (Fig. 8h). However, Fig. 8g also shows that HRRRx increases the number of underestimates to the west of the Sierra Nevada.

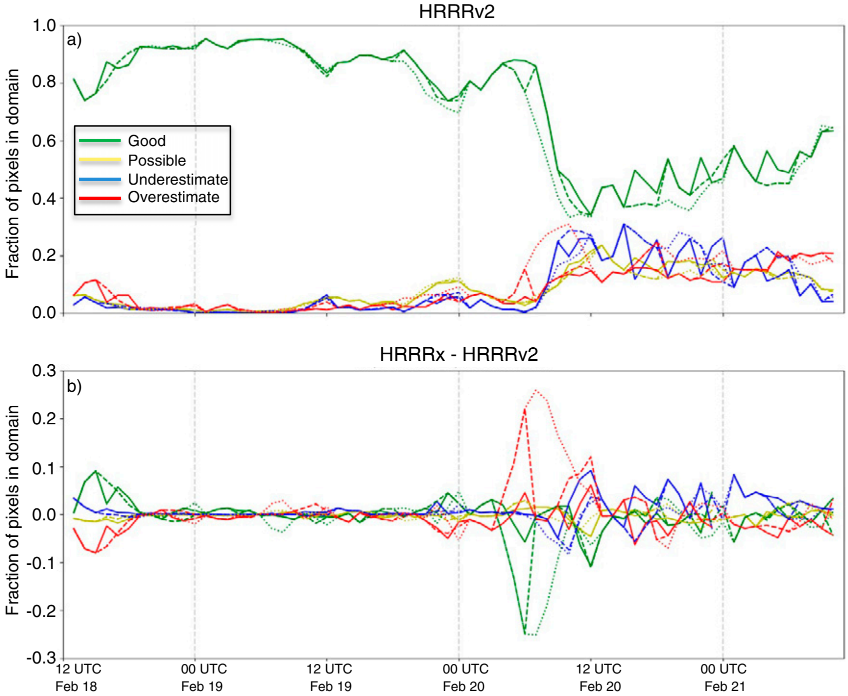

FIG. 6. (a) Time series of the fraction of pixels in the AQPI domain categorized as "good," "possible," "overestimated," and "underestimated" forecasts from the HRRRv2. (b) Time series difference in the percent of pixels in the AQPI domain falling into each of the forecast performance categories between the HRRRx and HRRRv2. Solid lines represent forecasts initialized every 3 h, dashed lines represent forecasts initialized every 6 h, and dotted lines represent forecasts initialized every 12 h. Dashed vertical gray lines indicate 0000 UTC each day for reference.

Next, we will demonstrate the evaluation methodology over longer-term aggregate statistics to track both the performance of an individual model version, and the differences between two versions of the model. The evaluation methodology was applied to seven AR events over the 2016–17 wet season with a total duration of 472 h.

The results from the 18–21 February case suggest that the HRRRx increases the number of underestimates in both the Central Valley and the western coastal mountains, while reducing the frequency of overestimation and increasing the number of good forecasts in the high elevations of the Sierra Nevada (although this result may be influenced by the presence of frozen precipitation). This pattern holds true for all of the events examined in this study in aggregate, as shown in Fig. 9 for the 3- and 12-h forecasts. The signal of reduced overestimates at high elevation is much more pronounced in the shorter forecast period, suggesting that data assimilation changes might be influencing the orographic precipitation (Bytheway et al. 2017). In fact, the longer forecast period shows a propensity for increasing overestimates at nearly all elevations.

Improvements at high-elevation are also seen in Fig. 10, which displays the difference between the HRRRx and HRRRv2 at the three forecast lengths binned by elevation. While the overall changes are relatively small (<5%), at elevations below 1000 m, there is a reduction in good forecasts at all forecast lengths, with a larger reduction for longer forecast periods. This mostly translates to an increase in overestimates in 6- and 12-h forecasts, while 3-h forecasts are dominated by an increased frequency of underestimation. At elevations greater than 1750 m, there is a very strong signal of the reduction of overestimates and increase in forecasts that are considered "good" for 3- and 6-h forecasts, while there is a slight increase in overestimates for 12-h forecasts at elevations up to 3000 m, similar to that seen in Fig. 9.

Figure 11 shows how grid boxes are categorized in HRRRx based on how they were categorized when evaluating the HRRRv2. At all forecast lengths, grid boxes categorized as "good" in HRRRv2 remain "good" nearly 95% of the time in the HRRRx (Fig. 11a). Additionally, grid boxes that were categorized as underestimates in HRRRv2 forecasts typically remain underestimates in HRRRx, with about 15% showing improvement (i.e., underestimates from HRRRv2 becoming "good" or "possible" forecasts) in the HRRRx (Fig. 11c). As suggested by the results shown in Figs. 9 and 10, the majority of the improvements made by the HRRRx were a result of reducing high biases. Grid boxes considered overestimates in the HRRRv2 were categorized as "possible" or "good" forecasts in the HRRRx nearly 60% of the time (Fig. 11d).

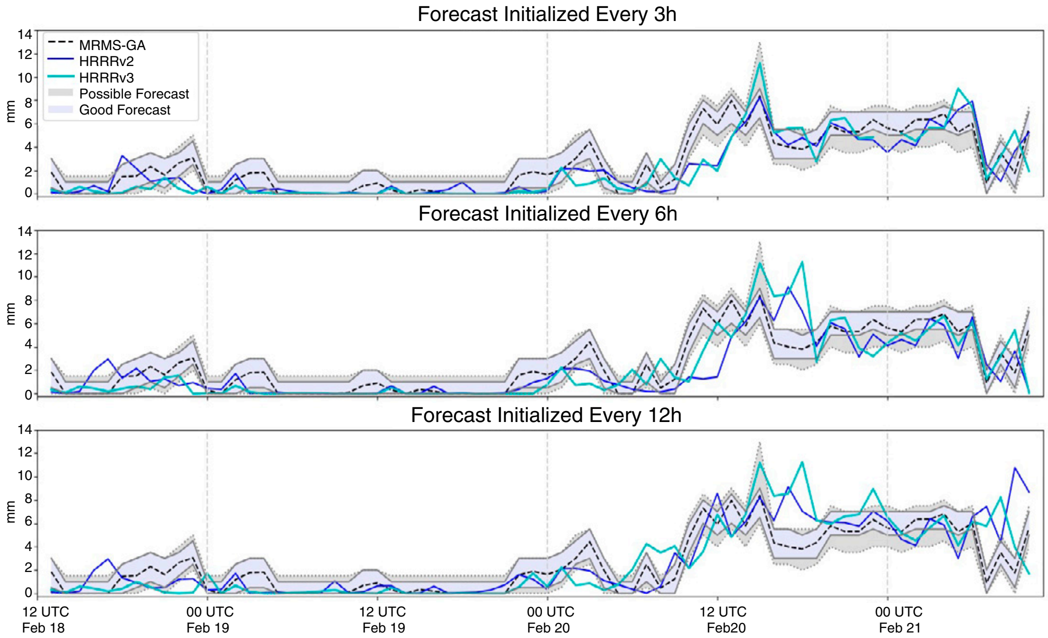

FIG. 7. As in Fig. 5, but with gauge datasets omitted from the QPE uncertainty calculations.

b. Effects of frozen precipitation

As noted above, frozen precipitation poses an additional challenge to accurate estimation of precipitation—many of the gauges in the HADS dataset do not measure frozen precipitation, and satellite-based QPE retrievals are known to struggle to both identify and quantify frozen precipitation. Additionally, previous studies have indicated that QPE datasets tend to exhibit a low bias at high elevations based on comparisons to snow datasets and that QPF may in fact outperform QPE under these conditions (Lundquist et al. 2019;

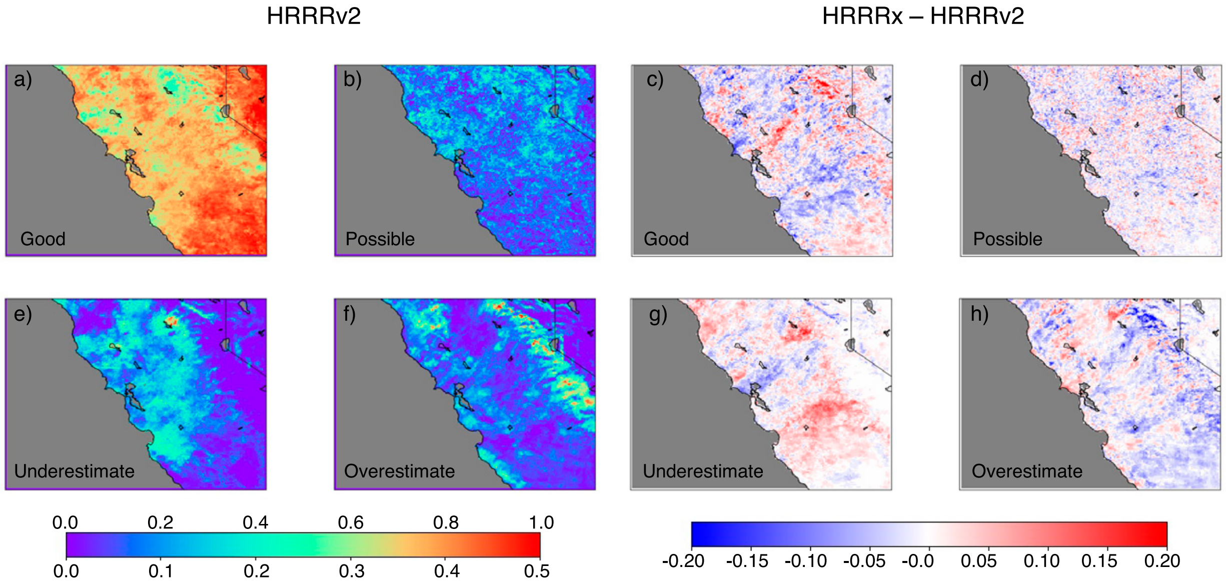

FIG. 8. (a),(b),(e),(f) Fraction of hours in the 18–21 Feb event that each HRRRv2 grid box is classified into each performance category for 3-h forecasts. (b),(c),(g),(h) Difference between HRRRx and HRRRv2 for the same event and forecast period. For HRRRv2, the top row of color bar values corresponds to "Good" forecasts in (a), while the lower row of values corresponds to (b), (e), and (f): "Possible," "Underestimates," and "Overestimates," respectively.

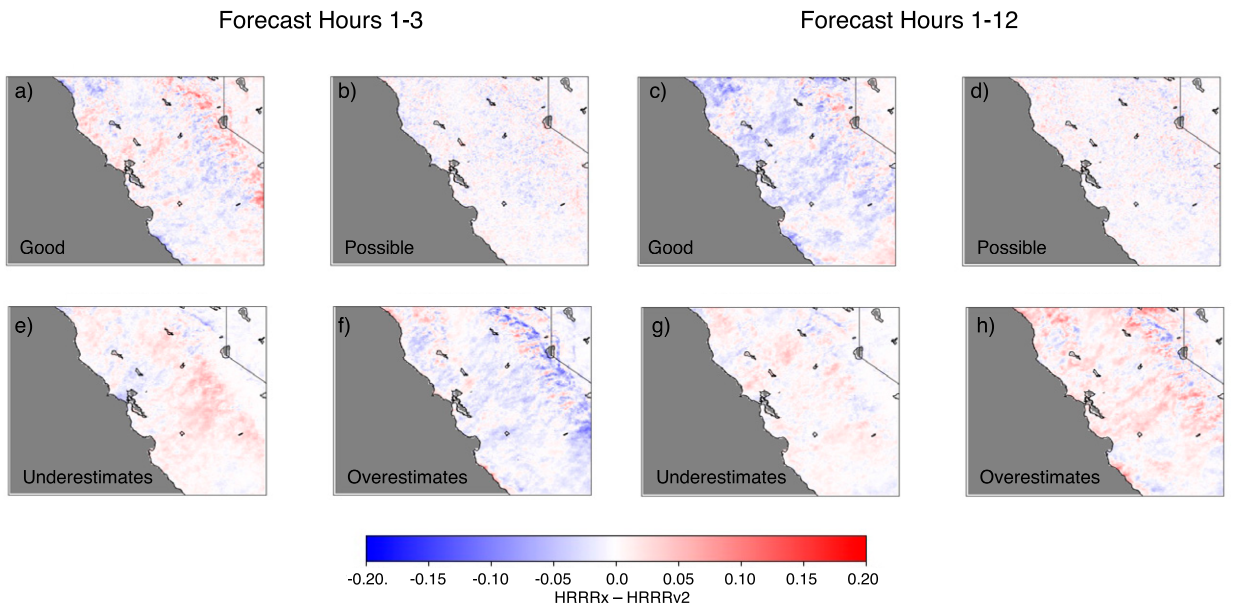

FIG. 9. As in Figs. 8c,d,g,h, aggregated over all seven AR events considered in this study for (a),(b),(e),(f) forecast hours 1–3 and (c),(d),(g),(h) forecast hours 1–12.

English et al. 2021b). As such, our assumption that "true" precipitation lies within the uncertainty of the various QPE is likely invalid under these conditions. To assess the impact of frozen precipitation on the evaluation, we repeated the assessment, using the HRRRv2 forecasts to eliminate hours and grid boxes when 2-m temperature fell below 273 K. We use only temperature forecasts from the HRRRv2 to ensure consistent temperatures are assigned to all datasets, and

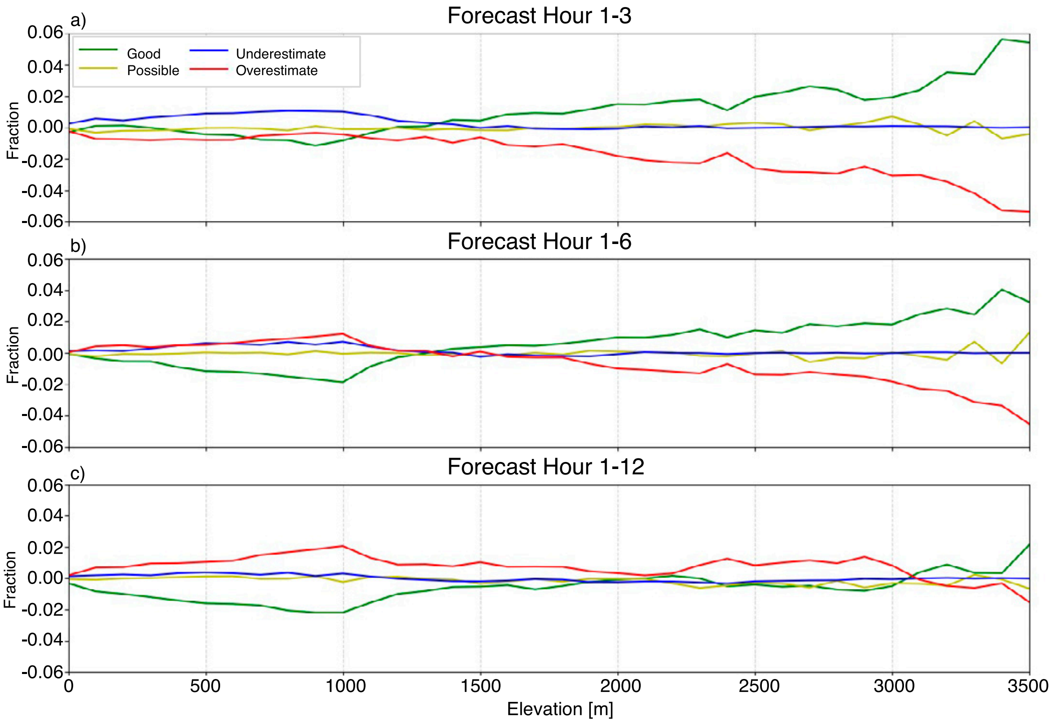

FIG. 10. Change in fraction of grid boxes falling into each evaluation category between HRRRx and HRRRv2 at each forecast length based on grid box elevation for all seven AR events included in this study. Dashed vertical gray lines indicate 500-m increments of elevation.

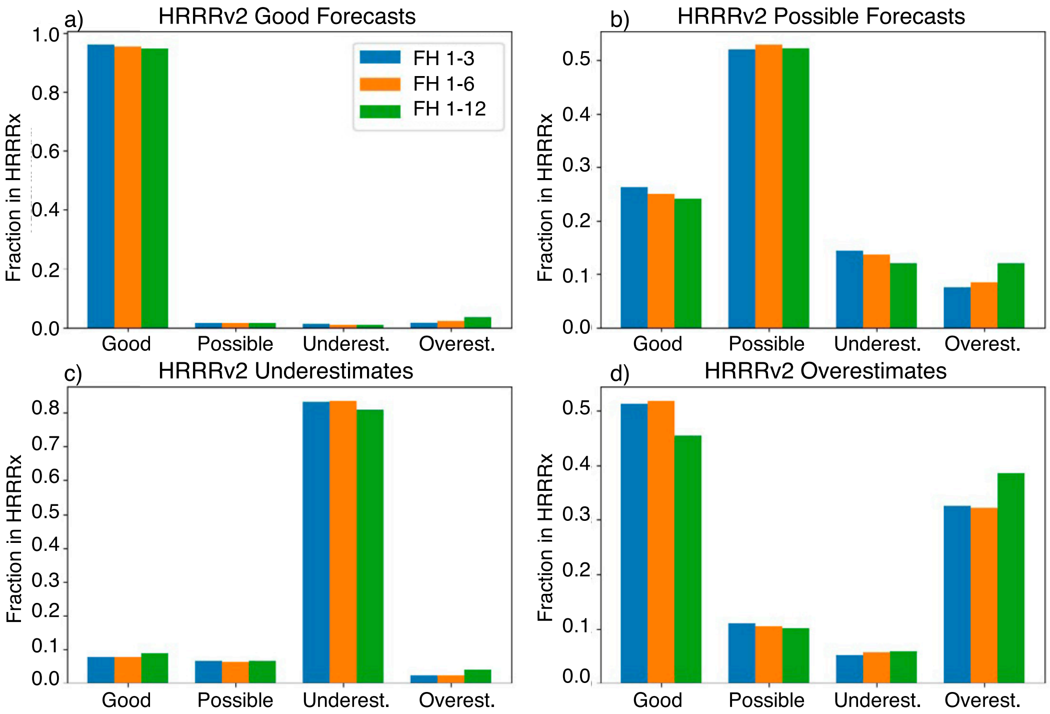

FIG. 11. HRRRx forecast performance categories as a function of HRRRv2 performance for grid boxes in HRRRv2 that were categorized as (a) "Good," (b) "Possible," (c) "Underestimates," and (d) "Overestimates" for all seven AR events included in this study. Blue bars indicate forecasts initialized every 3 h, orange bars show forecasts initialized every 6 h, and green bars represent forecasts initialized every 12 h.

avoid removing grid boxes where HRRRv2 and HRRRx may disagree on predicted temperature. Figure 12 shows the number of hours evaluated from 3-h forecasts over the course of the seven AR events (total duration = 472 h) when omitting instances of HRRRv2 predicted $T < 273$ K. Results were similar for the 6- and 12-h forecasts (not shown). Using the 273 K threshold, we are able to evaluate at least partial events over most of the domain, with the exception of the very highest elevations in the Sierra Nevada.

Excluding frozen precipitation from the 18–21 February case eliminates the highest terrain of the Sierra Nevada from evaluation, which can be seen in an area of gray on Fig. 13. In the Central Valley and most of the southern part of the domain, the results when excluding possible frozen precipitation remain unchanged, as do the fraction of hours in the event considered underestimates (Figs. 8e and 13e). There is some increase in the number of forecasts classified as possible surrounding the area that has been masked out. The largest change to the results from this event is a reduction in the fraction of forecasts falling into the "good" category on the windward side of the Sierra Nevada and the highest elevations of the Coast Range in the northern part of the domain (Figs. 8a and 13a). These changes correspond to an increase in overestimates in these areas (Figs. 8f and 13f). This is at least partly due to the smaller absolute numbers of hours considered, but also suggests that the 273-K temperature threshold may not capture all of the frozen precipitation in these areas. The improvement of HRRRx over HRRRv2 in these areas is slightly larger when frozen precipitation is removed (Figs. 13c,h).

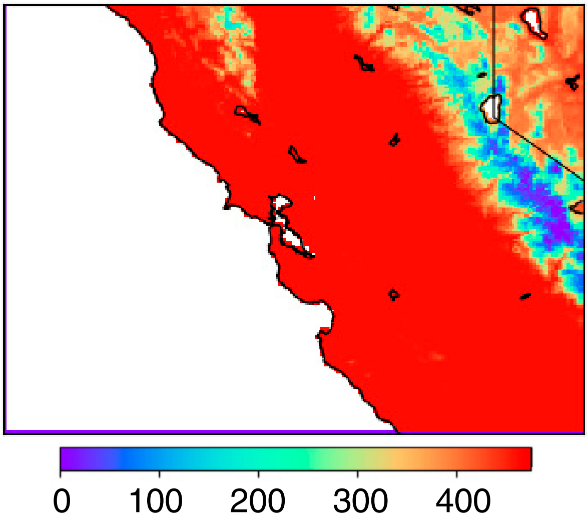

FIG. 12. Number of hours evaluated (out of a possible 472) when omitting forecast hours when HRRRv2 2-m temperature is below 273 K.

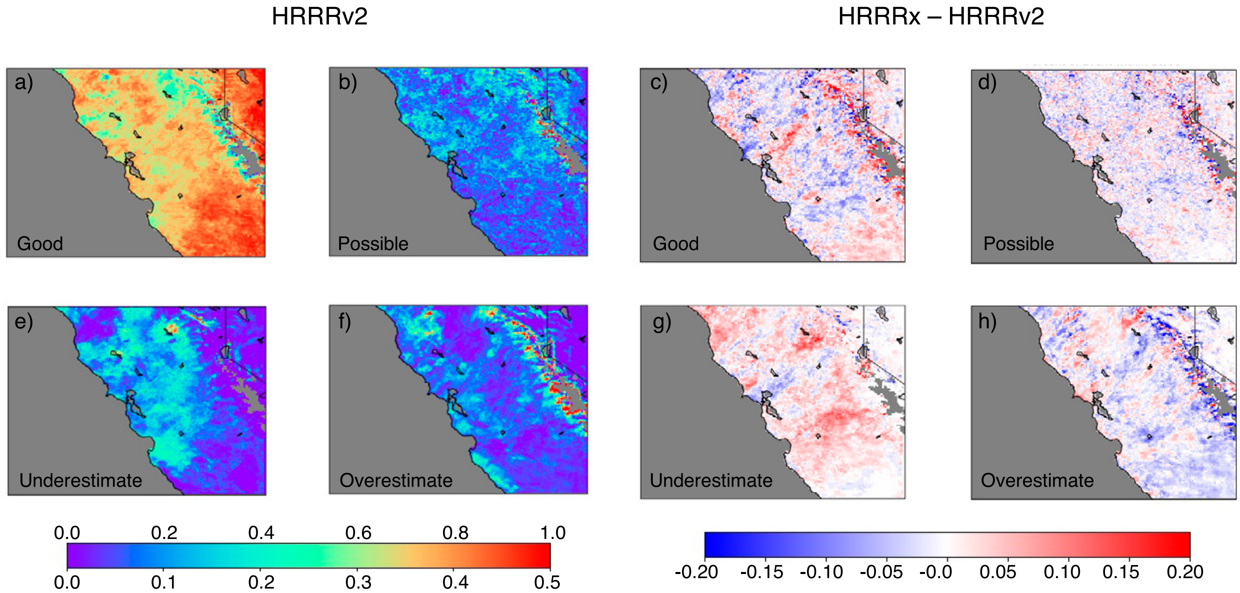

FIG. 13. As in Fig. 8, but for hours when HRRRv2 2-m temperature is greater than 273 K.

The results are similar when considering all seven AR events collectively. Figure 14 shows the same as Fig. 9 with frozen precipitation removed. For forecasts initialized every 3 h, there is a decrease in the fraction of improved forecasts on the windward side of the Sierra Nevada south of Lake Tahoe with corresponding increase in overestimates. However, there is also a large area in the highest terrain of the Sierra Nevada where the fraction of overestimates is reduced and the fraction of "good" forecasts is increased (Figs. 14a,f). Forecasts initialized every 12 h show similar patterns on the windward slopes south of Lake Tahoe; however, the corresponding improvement at the highest elevations is not present (Figs. 14c,h). This effect is well illustrated by Fig. 15, which repeats Fig. 10 with frozen precipitation removed. In Fig. 10 the HRRRx showed a sharp increase in the fraction of "good" forecasts with a corresponding decrease in

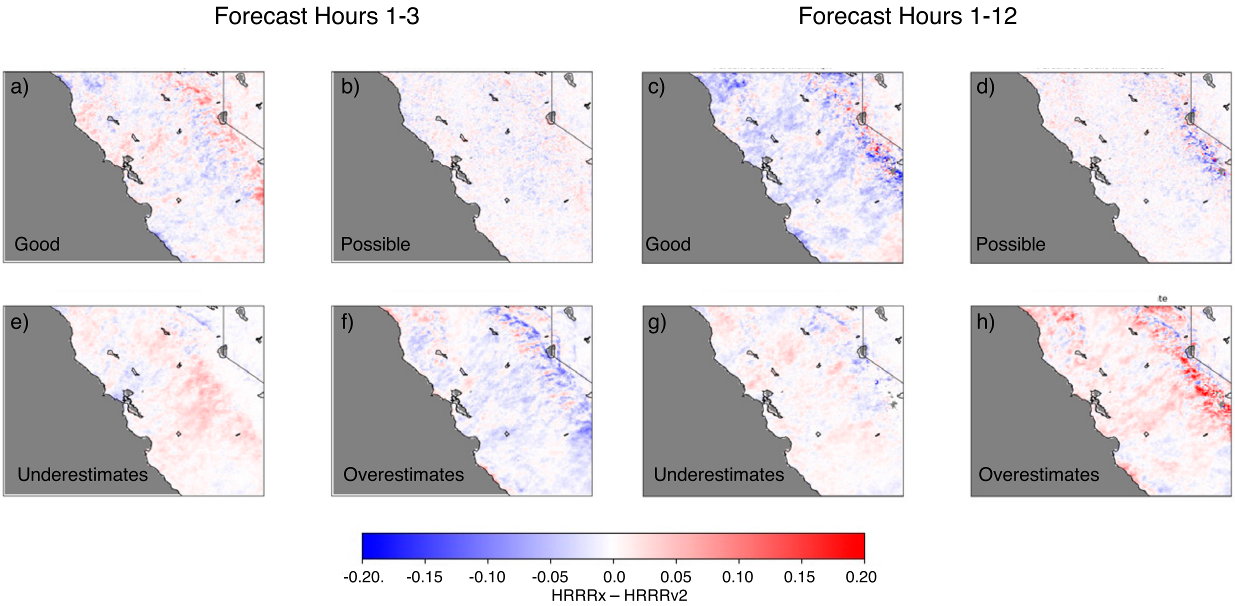

FIG. 14. As in Fig. 9, but for hours when HRRRv2 2-m temperature is greater than 273 K.

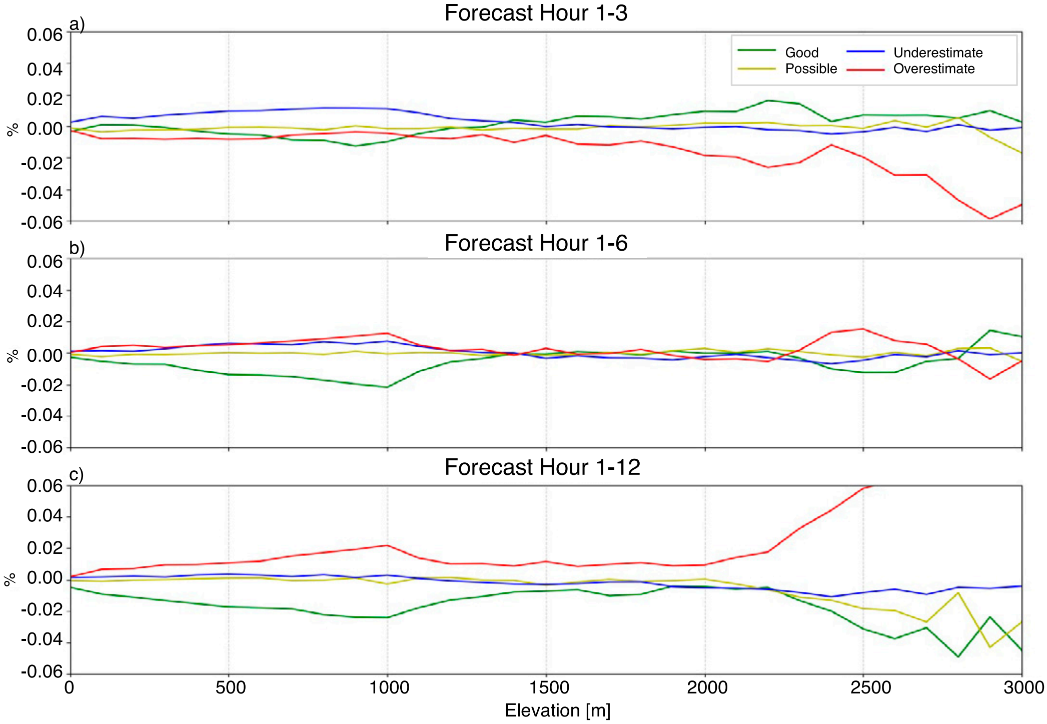

FIG. 15. As in Fig. 10, but for hours when HRRRv2 2-m temperature is greater than 273 K.

overestimated forecasts for elevations greater than 1500 m and forecasts initialized every 3 and 6 h. When considering only liquid precipitation, the increase in "good" forecasts is limited only to forecasts initialized every 3 h, and does not continue to increase at elevations greater than 2000 m. Similarly, the reduction in overestimates is limited to forecasts initialized every 3 h, and the pattern is strongly reversed for 12-h forecasts (i.e., overestimates increase and "good" forecasts decrease at elevations above 2000 m). As indicated by Fig. 12, no comparisons are performed at elevations greater than 3200 m, and results above 3000 m are not shown in Fig. 15 due to the very small number of available comparison points.

5. Discussion

a. Reference product selection

The pQPE methodology uses monthly CDFs of observed hourly precipitation in each grid box from each QPE product to describe the uncertainty of QPE estimates. The monthly time period was selected to balance the need for an adequate number of precipitating hours to produce the distributions and dampen the signal of any glaring outliers in the hourly estimates with the need to accurately represent the characteristics of the precipitation that fell, including extreme values. The evaluation methodology relies on the quantile of the median CDF for a given rain rate determined by a selected reference QPE product. For simplicity and demonstration purposes, we selected the MRMS-GA as our reference due to its performance when compared to gauges in the domain. The selection of a different reference product would produce somewhat different results, and a comparison of HRRR performance as a function of selected reference QPE could provide a demonstration of where the disagreement between QPE products is greatest. Indeed, it is likely that the "best" reference QPE product varies by location, season, or storm characteristics.

Selecting the QPE product with the CDF that most closely matches the monthly median CDF at each grid box would approximate a direct comparison with the median CDF and potentially provide a better initial guess of hourly precipitation. However, it would also add the complexity of using a location-dependent reference. Figure 16 shows which products' CDFs have the smallest RMSE compared to the monthly median in each 3-km grid box for January and February of 2017. While the MRMS-GA product compared most favorably to gauges based on several validation statistics, the IMERG monthly CDF most closely matches the all-QPE median CDF over a large portion of the domain in both months. Many areas are unchanged between the two months, but many areas would use a different reference QPE in January than in February. For example, in January, the RTMA CDF most closely matches the median in regions to the east and north of the San Francisco Bay, while in February these same areas are most closely represented by the URMA, MRMS-GA, and IMERG. An examination seeking common characteristics of areas where a given QPE product best matches the CDF median (e.g., terrain, gauge density, or precipitation characteristics) could provide insight into when and

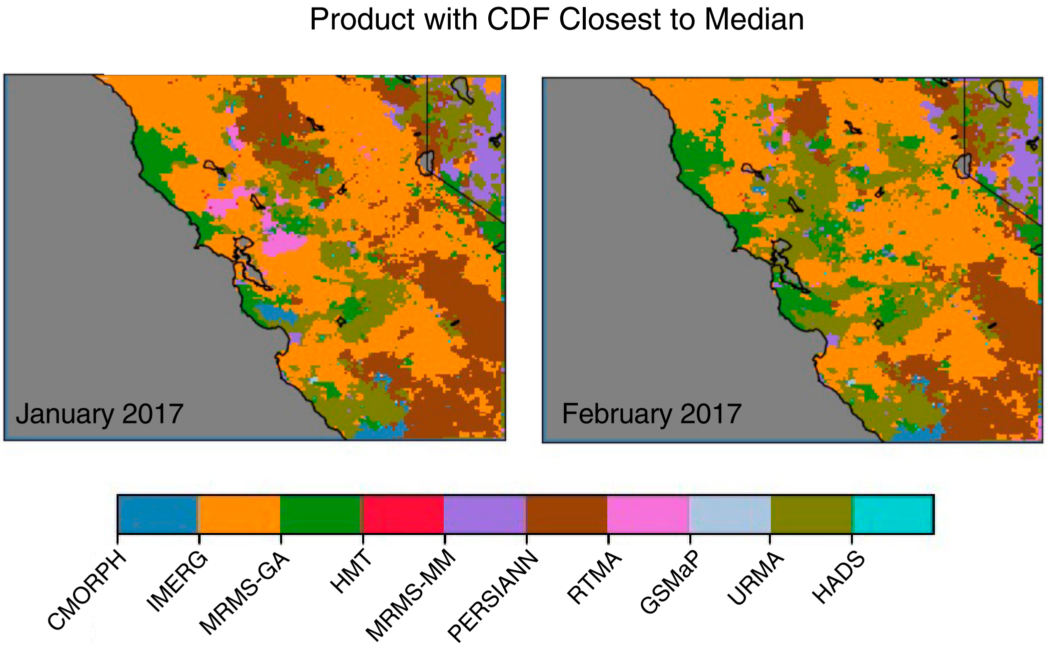

FIG. 16. QPE product with monthly CDF closest to the monthly median CDF (based on RMSE) for January and February 2017.

where an individual QPE might be considered more or less reliable, but is outside the scope of the current study.

b. *Testing the pQPE methodology through comparison to gauges*

The pQPE methodology relies on the assumption that the "true" rainfall lies somewhere within the range of the individual QPE estimates. The individual QPE products available in the AQPI domain each have their own strengths and weaknesses which contribute to the large QPE uncertainty in the region. The sometimes-large spread of QPE at a given location and time increases the likelihood that the true rainfall is found within the range of the estimates. In spite of this, it is a worthwhile exercise to examine how well the two gauge-only QPE products relate to the range of "good" and "possible" precipitation estimates provided by the pQPE methodology, since surface rain gauges are typically considered to be the most reliable estimate of the true rainfall. This test is of course limited only to grid boxes that have a gauge available, and also comes with the caveat that the gauge is a point measurement, which may not represent the areal average rainfall in the grid box, particularly in the complex terrain.

Over the course of the 7 AR events, there are a total of 92 976 valid HADS observations, and 20 955 HMT observations. Of these observations 13.2% (7.2%) of the HADS precipitation estimates were outside the IQR (10th/90th percentiles). The HMT gauges compared slightly worse, with 21% (13.1%) of these estimates outside of the IQR (10th/90th percentile). Of the HADS observations that were outside the IQR, 17.4% occurred during periods when the forecast temperature was less than 273 K (9.2% for the HMT). Figure 17 shows the fraction of hours during the seven AR events that individual gauges were outside the IQR and 10th/90th percentiles. The majority of the gauges outside the IQR for both networks are located in the complex terrain along the coastal mountains and Sierra Nevada, and these maps imply that a relatively small number of individual gauges are contributing a large fraction of the hours that these products disagree with the range of "good" and "possible" hourly rainfall values. For example, five gauges in the HMT network are outside the IQR 35% of the time, corresponding to 18% of the observations falling outside the IQR of the QPE products. All of the gauges showing the largest fraction of hours outside the IQR show a reduced fraction of hours outside the 10th/90th percentiles. Again, the most frequent disagreements are mainly located in the terrain, and are most often the result of the gauges indicating more precipitation than the remainder of the QPE products. If the gauges are believed, it is possible that many of the HRRRv2 forecasts classified as overestimates in the complex terrain are more reliable than indicated by this methodology. However, there are several possible reasons for the disagreement between gauges and the pQPE, including gauge overestimates due to a buildup of snow that melts rapidly, the possibility of frozen precipitation effecting the gridded QPE, gauges, or both, or, in areas where the gauges are too low, gauge undercatch due to wind,. Alternatively, the gridded products could all systematically over or underestimate precipitation for reasons noted in the introduction.

c. *Sample size limitations*

Due to the relatively infrequent nature of atmospheric river events and incomplete HRRRx archive, this study

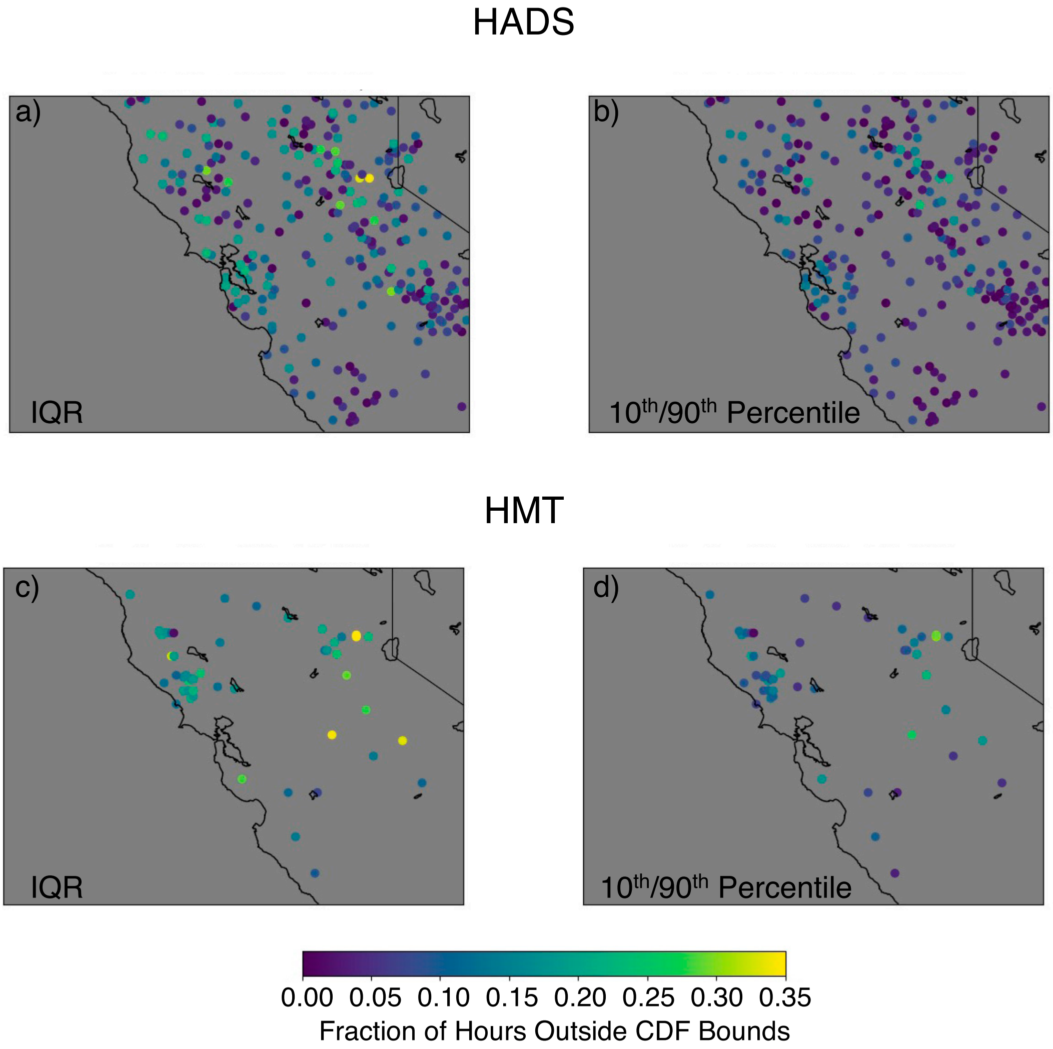

FIG. 17. Fraction of hours when (top) HADS and (bottom) HMT gauge observations lie outside the (a),(c) IQR and (b),(d) 10th and 90th percentiles of the all-QPE CDFs for all hours of the seven atmospheric river events.

consisted of only a small number of cases, and we emphasize that the comparisons shown herein are meant only to demonstrate the pQPE methodology. A much larger data sample would be required to make robust generalizations about performance differences between the versions of the HRRR. Additionally, HRRRx was under development during the period of study, and therefore it is possible that there were small differences in the model configuration during the different AR events. For the purpose of the AQPI project, forecast model experiments are typically run on a similarly small number of cases, and this QPF evaluation methodology can quickly show the changes in forecast precipitation that result from experimental changes to the model. Results from individual cases can be useful to examine how changes to the model influence forecast performance as a function of storm characteristics, such as the angle of incidence of the winds to the terrain.

5. Summary and conclusions

In this paper, we present a methodology to make use of all available high resolution (hourly, <10 km) QPE information in the AQPI domain in order to evaluate high resolution forecast model performance. The use of a large number of QPE products represents an attempt to account for the large uncertainty in high resolution QPE in the region. Considering the estimates from all available high resolution QPE products reduces the chance of penalizing the model performance metrics when a given reference QPE is unreliable, for example when radar beams are blocked or gauges are clogged, without requiring prior knowledge of the potential issue in the observed data. It is worth noting that there may be some situations where all available QPE are unreliable, for example frozen precipitation over snow cover in a region of radar beam blockage.

The results of the demonstration given here show the applicability of the pQPE evaluation methodology to assess model performance of individual forecast periods, precipitation events, or to evaluate changes between versions of the same model over extended periods of time. The example comparison between HRRRv2 and HRRRx suggests that the HRRRx reduces the tendency of the HRRRv2 to overpredict precipitation in the high terrain of the Sierra Nevada, particularly at short forecast lengths. However, because some studies have shown that QPE datasets tend to be biased low at high elevations based on comparisons to snow datasets (Lundquist et al. 2019; English et al. 2021b), we repeated the evaluation while omitting frozen precipitation based on 2-m temperature from the HRRRv2. This eliminated grid boxes above ∼3200-m elevation from consideration, and reduced the number of hours evaluated in the Sierra Nevada and the highest elevations of the Coastal Range, therefore reducing the apparent improvements made in these regions by the HRRRx. While many of the gauges included in the QPE products used in this evaluation do not measure frozen precipitation, many of them are collocated with snow measurements, which could be included as another source of QPE.

The development of a method to account for the large QPE uncertainty when validating high resolution QPF should not serve as a substitute for continued work to improve high resolution QPE, especially in complex terrain such as in the AQPI domain where frozen precipitation remains a challenge, even for the described pQPE methodology. Continued work to add instrumentation, improve observations of frozen precipitation, merge different observations into multisensor QPE products, and to understand the physical characteristics of storms that may lead to higher or lower uncertainty in precipitation estimates will be crucial to our ability to evaluate future forecast models and monitor and manage water resources in complex terrain regions.

Acknowledgments. This study was supported by Grant 4600011485, funded by the Safe Drinking Water, Water Quality and Supply, Flood Control, River and Coastal Protection Bond Act of 2006, administered by the State of California, Department of Water Resources, the NOAA Physical Sciences Laboratory (PSL), and the NOAA Cooperative Agreement with CIRES, NA17OAR4320101. We are grateful to PSL staff for maintaining the HMT network and making the data available as well as to all the organizations and individuals that provided data for this study, including Ying Lin for access to archived URMA. Finally, we thank Curtis Alexander and Dave Turner from the Global Systems Laboratory for making the HRRR data available and for providing useful feedback as part of this study.

Data availability statement. Access to publicly available datasets is outlined in individual data descriptions in section 2, except RTMA, which is archived in the National Digital Guidance Database (NDGD) at the National Centers for Environmental Information (NCEI). 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. HRRR operational output is also archived at NCEP as well as Google Cloud at https://console.cloud.google.com/marketplace/product/noaa-public/hrrr.

REFERENCES

Bartsotas, N. S., E. N. Anagnostou, E. I. Nikolopoulos, and G. Kallos, 2018: Investigating satellite precipitation uncertainty over complex terrain. *J. Geophys. Res. Atmos.*, **123**, 5346–5359, https://doi.org/10.1029/2017JD027559.

Beck, H. E., and Coauthors, 2019: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrol. Earth Syst. Sci.*, **23**, 207–224, https://doi.org/10.5194/hess-23-207-2019.

Bellerby, T. J., and J. Sun, 2005: Probabilistic and ensemble representations of the uncertainty in an IR/Microwave satellite precipitation product. *J. Hydrometeor.*, **6**, 1032–1044, https://doi.org/10.1175/JHM454.1.

Benjamin, S., and Coauthors, 2016: A North American hour 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., and C. D. Kummerow, 2015: Toward an object-based assessment of high-resolution forecasts of long-lived convective precipitation in the central U.S. *J. Adv. Model. Earth Syst.*, **7**, 1248–1264, https://doi.org/10.1002/2015MS000497.

——, ——, and C. Alexander, 2017: A features-based assessment of the evolution of warm season precipitation forecasts from the HRRR model over three years of development. *Wea. Forecasting*, **32**, 1841–1856, https://doi.org/10.1175/WAF-D-17-0050.1.

——, M. Hughes, K. Mahoney, and R. Cifelli, 2019: A multiscale evaluation of multisensor quantitative precipitation estimates in the Russian River Basin. *J. Hydrometeor.*, **20**, 447–466, https://doi.org/10.1175/JHM-D-18-0142.1.

——, ——, ——, and ——, 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.

Cai, H., and R. E. Dumais Jr., 2015: Object-based evaluation of a numerical weather prediction model's performance through forecast storm characteristic analysis. *Wea. Forecasting*, **30**, 1451–1468, https://doi.org/10.1175/WAF-D-15-0008.1.

Ciach, G. J., W. F. Krajewski, and G. Villarini., 2007: Product-error-driven uncertainty model for probabilistic quantitative precipitation estimation with NEXRAD data. *J. Hydrometeor.*, **8**, 1325–1347, https://doi.org/10.1175/2007JHM814.1.

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., R. G. Bullock, T. L. Jensen, M. Xue, and F. Kong, 2014: Application of object-based time-domain diagnostics for tracking precipitation systems in convection-allowing models. *Wea. Forecasting*, **29**, 517–542, https://doi.org/10.1175/WAF-D-13-00098.1.

Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.*, **33**, 140–158, https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

——, M. E. Slater, J. A. Roberti, S. H. Saseter, and L. W. Swift Jr., 2017: High-resolution precipitation mapping in a mountainous watershed: Ground truth for evaluating uncertainty in a national precipitation dataset. *Int. J. Climatol.*, **37**, 124–137, https://doi.org/10.1002/joc.4986.

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, 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.

De Pondeca, M. S. F. V., and Coauthors, 2011: The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: Current status and development. *Wea. Forecasting*, **26**, 593–612, https://doi.org/10.1175/WAF-D-10-05037.1.

Derin, Y., and Coauthors, 2016: Multiregional satellite precipitation products evaluation over complex terrain. *J. Hydrometeor.*, **17**, 1817–1836, https://doi.org/10.1175/JHM-D-15-0197.1.

Dinku, T., S. Chidzambwa, P. Ceccato, S. J. Connor, and C. F. Ropelewski, 2008: Validation of high-resolution satellite rainfall products over complex terrain. *Int. J. Remote Sens.*, **29**, 4097–4110, https://doi.org/10.1080/01431160701772526.

——, F. Ruiz, S. J. Connor, and P. Ceccato, 2010: Validation and intercomparison of satellite rainfall estimates over Columbia. *J. Appl. Meteor. Climatol.*, **49**, 1004–1014, https://doi.org/10.1175/2009JAMC2260.1.

Dougherty, K. J., J. D. Horel, and J. E. Nachamkin, 2021: Forecast skill for California heavy precipitation periods from the High-Resolution Rapid Refresh model and the Coupled Ocean–Atmosphere Mesoscale Prediction System. *Wea. Forecasting*, https://doi.org/10.1175/WAF-D-20-0182.1, in press.

Ebert, E. E., and W. A. Gallus Jr., 2009: Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. *Wea. Forecasting*, **24**, 1401–1415, https://doi.org/10.1175/2009WAF2222252.1.

——, J. E. Janowiak, and C. Kidd, 2007: Comparison of near-real-time precipitation estimates from satellite observations and numerical models. *Bull. Amer. Meteor. Soc.*, **88**, 47–64, https://doi.org/10.1175/BAMS-88-1-47.

English, J. M., D. D. Turner, T. I. Alcott, W. R. Moninger, J. L. Bytheway, and R. Cifelli, 2021a: AQPI: RAP/HRRR model forecasts of California atmospheric river events. *35th Conf. on Hydrology,* Virtual, Amer. Meteor. Soc., 931, https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Paper/379010.

——, ——, ——, ——, ——, ——, and M. Marquis, 2021b: Evaluating operational and experimental HRRR model forecasts of atmospheric river events in California. *Wea. Forecasting*, **36**, 1925–1944, https://doi.org/10.1175/WAF-D-21-0081.1.

Gourley, J. J., Y. Hong, Z. G. Flamig, L. Li, and J. Wang, 2010: Intercomparison of rainfall estimates from radar, satellite, gauge, and combinations for a season of record rainfall. *J. Appl. Meteor. Climatol.*, **49**, 437–452, https://doi.org/10.1175/2009JAMC2302.1.

Hirpa, F. A., M. Gebremichael, and T. Hopson, 2010: Evaluation of high-resolution satellite precipitation products over very complex terrain in Ethiopia. *J. Appl. Meteor. Climatol.*, **49**, 1044–1051, https://doi.org/10.1175/2009JAMC2298.1.

Hong, Y., K. Hsu, S. Sorooshian, and X. Gao, 2004: Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. *J. Appl. Meteor.*, **43**, 1834–1853, https://doi.org/10.1175/JAM2173.1.

Hsu, K., X. Gao, ——, and H. V. Gupta, 1997: Precipitation estimation from remotely sensed information using artificial neural networks. *J. Appl. Meteor.*, **36**, 1176–1190, https://doi.org/10.1175/1520-0450(1997)036<1176:PEFRSI>2.0.CO;2.

Huffman, G. J., and Coauthors, 2018: NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG). Algorithm Theoretical Basis Doc., version 5.2, 35 pp., https://pmm.nasa.gov/sites/default/files/document_files/IMERG_ATBD_V5.2_0.pdf.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. Xie, 2004: CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydrometeor.*, **5**, 487–503, https://doi.org/10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2.

Kim, D., B. Nelson, and D.-J. 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.

Kirstetter, P.-E., Y. Hong, J. J. Gourley, M. Schwaller, W. Petersen, and Q. Cao, 2015: Impact of sub-pixel rainfall variability on spaceborne precipitation estimation: Evaluating the TRMM 2A25 product. *Quart. J. Roy. Meteor. Soc.*, **141**, 953–966, https://doi.org/10.1002/qj.2416.

Kubota, T., and Coauthors, 2007: Global precipitation map using satelliteborne microwave radiometers by the GSMaP project: Production and validation. *IEEE Trans. Geosci. Remote Sens.*, **45**, 2259–2275, https://doi.org/10.1109/TGRS.2007.895337.

Lin, Y., and K. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2, https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.

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.

Maddox, R. A., J. Zhang, J. J. Gourley, and K. W. Howard, 2002: Weather radar coverage over the contiguous United States. *Wea. Forecasting*, **17**, 927–934, https://doi.org/10.1175/1520-0434(2002)017<0927:WRCOTC>2.0.CO;2.

Nelson, B. R., O. P. Prat, D.-J. 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.

Ryzhkov, A. V., T. J. Schuur, D. W. Burgess, P. L. Heinselman, S. E. Giangrande, and D. S. Zrnic, 2005: The joint polarization experiment: Polarimetric rainfall measurements and hydrometeor classification. *Bull. Amer. Meteor. Soc.*, **86**, 809–824, https://doi.org/10.1175/BAMS-86-6-809.

Sun, Q., C. Maio, Q. Duan, H. Ashouri, S. Sorooshian, and K.-L. Hsu, 2018: A review of global precipitation data set: Data sources, estimation and intercomparisons. *Rev. Geophys.*, **56**, 79–107, https://doi.org/10.1002/2017RG000574.

Swain, D. L., B. Langenbrunner, J. D. Neelin, and A. Hall, 2018: Increasing precipitation volatility in twenty-first-century California. *Nat. Climate Change*, **8**, 427–433, https://doi.org/10.1038/s41558-018-0140-y.

Tan, J., G. J. Huffman, D. T. Bolvin, and E. J. Nelkin, 2019: IMERG V06: Changes to the morphing algorithm. *J. Atmos. Oceanic Technol.*, **36**, 2471–2482, https://doi.org/10.1175/JTECH-D-19-0114.1.

Tian, Y., and C. D. Peters-Lidard, 2010: A global map of uncertainties in satellite-based precipitation measurements. *Geophys. Res. Lett.*, **37**, L24407, https://doi.org/10.1029/2010GL046008.

Timmermans, B., M. Wehner, D. Cooley, T. O'Brien, and H. Krishnan, 2019: An evaluation of the consistency of extremes in gridded precipitation data sets. *Climate Dyn.*, **52**, 6651–6670, https://doi.org/10.1007/s00382-018-4537-0.

Ushio, T., and Coauthors, 2009: A Kalman filter approach to the Global Satellite Mapping of Precipitation (GSMaP) from combined passive microwave and infrared radiometric data. *J. Meteor. Soc. Japan*, **87A**, 137–151, https://doi.org/10.2151/jmsj.87A.137.

Villarini, G., and W. F. Krajewski, 2009: Review of the different sources of uncertainty in single polarization radar-based estimates of rainfall. *Surv. Geophys.*, **31**, 107–129, https://doi.org/10.1007/s10712-009-9079-x.

——, ——, G. J. Ciach, and D. L. Zimmerman, 2009a: Product-error-driven generator of probable rainfall conditioned on WSR-88D precipitation estimates. *Water Resour. Res.*, **45**, W01404, https://doi.org/10.1029/2008WR006946.

——, ——, and J. A. Smith, 2009b: New paradigm for statistical validation of satellite precipitation estimates: Application to a large sample of the TMPA 0.25° 3-hourly estimates over Oklahoma. *J. Geophys. Res.*, **114**, D12106, https://doi.org/10.1029/2008JD011475.

Willie, D., H. Chen, V. Chandrasekar, R. Cifelli, C. Campbell, D. Reynolds, S. Matrosov, and Y. Zhang, 2017: Evaluation of multisensor quantitative precipitation estimation in Russian River basin. *J. Hydrol. Eng.*, **22**, E5016002, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001422.

Xie, P., R. Joyce, S. Wu, S.-H. Yoo, Y. Yarosh, F. Sun, and R. Lin, 2017: Reprocessed, bias-corrected CMORPH global high resolution precipitation estimates from 1998. *J. Hydrometeor.*, **18**, 1617–1641, https://doi.org/10.1175/JHM-D-16-0168.1.

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.

——, Y. Qi, C. Langston, B. Kaney, and K. Howard, 2014: A real-time algorithm for merging radar QPEs with rain gauge observations and orographic precipitation climatology. *J. Hydrometeor.*, **15**, 1794–1809, https://doi.org/10.1175/JHM-D-13-0163.1.

——, 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.