

# An Evaluation of Integrated Water Vapor, Wind, and Precipitation Forecasts Using Water Vapor Flux Observations in the Western United States

LISA S. DARBY, ALLEN B. WHITE, AND DANIEL J. GOTTAS

*NOAA/Earth System Research Laboratory, Boulder, Colorado*

TIMOTHY COLEMAN

*University of Colorado Boulder, Boulder, Colorado*

(Manuscript received 14 September 2018, in final form 5 September 2019)

## ABSTRACT

Differences between forecasts and observations at eight atmospheric river observatories (AROs) in the western United States during winter 2015/16 are analyzed. NOAA's operational RAP and HRRR 3-h forecasts of wind, integrated water vapor (IWV), integrated water vapor flux (IWV flux), and precipitation from the grid points nearest the AROs were paired with ARO observations presented in the NOAA/Physical Sciences Division's water vapor flux tool (WVFT). The focus of this paper is to characterize and quantify the differences in the WVFT observations and forecasts. We used traditional forecast evaluation methods since they were compatible with the design of the tool: a near-real-time visual depiction of hourly observed and forecasted variables at a single location. Forecast root-mean-squared errors (RMSEs) and unbiased RMSEs, standard deviations of the observed and forecasted variables, and frequency bias scores (FBS) for all of the fields, plus equitable threat scores for precipitation, are presented. Both models forecasted IWV at all AROs and the winds that drive orographic precipitation at most AROs within a reasonable range of the observations as indicated by comparisons of the standard deviations and RMSEs of the forecasts with the standard deviations of the observations and FBS. These results indicated that forecasted advection of moisture to the stations was adequate for generating precipitation. At most stations and most hourly precipitation rates, the HRRR underpredicted precipitation. At several AROs the RAP precipitation forecasts more closely matched the observations at smaller ( $<1.27 \text{ mm h}^{-1}$ ) precipitation rates, but underpredicted precipitation rates  $> 2 \text{ mm h}^{-1}$ .

## 1. Introduction

Since 2007, the Physical Sciences Division (PSD) of the National Oceanic and Atmospheric Administration's (NOAA) Earth System Research Laboratory (ESRL) has provided an online water vapor flux tool (WVFT), with observations and forecasts available in near-real time (<https://www.esrl.noaa.gov/psd/data/obs/datadisplay/>). The WVFT displays of measurements and 3-h forecasts are useful for predicting precipitation at several locations in the western United States, indicating when these locations are at risk for heavy precipitation. The tool also provides a quick visual check of whether the models have been accurately predicting precipitation, winds, and moisture advection during the last 48 h at each site. The WVFT is a product of years of PSD precipitation

observations and research related to the midlatitude atmospheric dynamics and physical processes that produce extreme precipitation events. The scientific foundation for the development of the tool is based on results from key papers from several major field projects, including the California Land-falling Jets Experiment (CALJET; Ralph et al. 1999, 2003, 2004; Neiman et al. 2002), the Pacific Land-falling Jets Experiments (PACJET 2001–03; Neiman et al. 2005, 2006; Ralph et al. 2005), and the Hydrometeorology Testbed–West (HMT–West; Ralph et al. 2010; White et al. 2013, 2015). Neiman et al. (2009) and White et al. (2012, 2013) explain the details of the measurements at atmospheric river observatories (ARO) where the measurements that support the WVFT occur, and the scientific background behind the development of the tool. The observations analyzed in this paper were originally developed and supported by the NOAA Hydrometeorology Testbed (Ralph et al. 2013), and, as a

*Corresponding author:* Lisa S. Darby, [lisa.darby@noaa.gov](mailto:lisa.darby@noaa.gov)

DOI: 10.1175/WAF-D-18-0159.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](https://www.ametsoc.org/PUBSReuseLicenses) ([www.ametsoc.org/PUBSReuseLicenses](https://www.ametsoc.org/PUBSReuseLicenses)).

legacy of HMT-West and other work (Ralph et al. 2014), PSD and partners continue to maintain AROs, and thus the WVFT, in the western United States.

Originally, the WVFT was a web-based situational awareness and diagnostic tool that included a 24-h time series of hourly averaged near-real-time observations and thresholds of key variables indicating when observed conditions were favorable for potential extreme precipitation (Neiman et al. 2009). The tool has since evolved to include 48 h of observations and thresholds, and, beginning in 2014, 3-h forecasts of the same meteorological variables shown in the observation portion of the WVFT. The observations, thresholds, and forecasts are all components of the WVFT. The WVFT updates hourly in near-real time, providing a concise display for forecasters to assess low-level advection of column-integrated water vapor toward mountainous terrain that may produce orographic precipitation (Neiman et al. 2009).

The observations are ingested into the WVFT from the collocated instruments that comprise an ARO (White et al. 2012, 2013, 2015): radar wind profiler, snow-level radar, GPS for integrated water vapor (IWV), and standard surface meteorology measurements. Observations displayed on the WVFT include wind profiles, snow level, upslope wind speed (derived from the wind profiles), IWV, integrated water vapor flux (a product of the upslope wind speed and the IWV), and precipitation (Fig. 1).

Twelve hours of forecasts of the same meteorological variables shown in the observation portion of the WVFT, with a 3-h lead time, are appended to immediately follow the last hourly observation (designated by a vertical line in the wind profile panel). Note that following wind profiler convention, time increases from right to left along the  $x$  axis. This arrangement allows the forecaster to track the “advection of coherent atmospheric features from west to east in a traditional meteorological framework” (Neiman et al. 2009) and to quickly determine in near-real time if the 3-h forecasts have been consistent with the observations. Forecasts are available from these NCEP operational forecast models: 1) the Rapid Refresh (RAP) and 2) the High-Resolution Rapid Refresh (HRRR), (Benjamin et al. 2016a,b) and in some cases, the experimental versions of these models. Links to each model forecast appear in the upper-left corner of the WVFT web page for each ARO so the user can select a model to view.

Superimposed upon the observations and forecasts are the minimum thresholds of upslope wind speed and direction, IWV, and IWV flux found to be needed for extreme precipitation (Neiman et al. 2009). The thresholds were originally determined for Bodega Bay (BBY; Fig. 2)

and in winter 2015/16 all stations used the BBY thresholds. To develop these thresholds, Neiman et al. (2009) examined precipitation events observed over four winter seasons (WYs 2001, 2004, 2005, and 2006) by the ARO at BBY and at a downwind coastal mountain site [Cazadero (CZC); 475-m elevation] and found that the heavier ( $>10 \text{ mm h}^{-1}$ ) orographic precipitation events occurred almost exclusively when the upslope component of the wind between 850 and 1150 m AGL exceeded  $12.5 \text{ m s}^{-1}$  and the IWV exceeded 2 cm (relating to an IWV flux threshold of  $25 \text{ cm m s}^{-1}$ ). The upslope layer is defined as the 300-m layer of wind profiler observations surrounding the 1-km level where the correlation between upslope wind and orographic precipitation is maximized (Neiman et al. 2009). This  $12.5 \text{ m s}^{-1}$  wind threshold is consistent with Waliser and Guan's (2017) findings that surface winds of  $12 \text{ m s}^{-1}$  are associated with atmospheric river events around the globe, including the west coast of the United States. It is these thresholds that provide a means to diagnose the orographic forcing components of precipitation and are a value-added aspect of the WVFT.

Over the years, visual inspection of the WVFT revealed discrepancies, sometimes quite large, between the 3-h precipitation forecasts and the observations, yet the forecasts of winds, IWV, and IWV flux appeared to more closely match the observations much of the time. The focus of this paper is to evaluate the WVFT forecasts at eight sites (Fig. 2, Table 1) that had all the requisite instrumentation to produce the tool during winter 2015/16. We quantify the precipitation forecast errors for each site, as well as the “ingredients” of the orographic precipitation—upslope winds and moisture. The WVFT forecasts have not previously been evaluated in this way.

Five of the ARO sites during our study period of winter 2015/16 were near the coast (Fig. 2): Forks, Washington (FKS); Astoria, Oregon (AST); Northbend, Oregon (OTH); Cazadero, California (CZC); and Bodega Bay, California (BBY). Troutdale, Oregon (TDE), was  $\sim 160$  km inland. Wasco, Oregon (WCO), and Chico, California (CCO), were inland stations that had mountain ranges between them and the coast.

Evaluations of HRRR and RAP precipitation forecasts found in the current literature tend to focus on either wintertime precipitation type, such as evaluations of model forecasts of rain, snow, and mixed-phase precipitation (Ikeda et al. 2013; Benjamin et al. 2016a; Elmore et al. 2015) or the characteristics of warm season convection (Pinto et al. 2015; Hwang et al. 2015; Cai and Dumais 2015), rather than quantitative precipitation forecasts (QPF). Studies that include evaluations of HRRR or RAP QPF include Peckham et al. (2016),

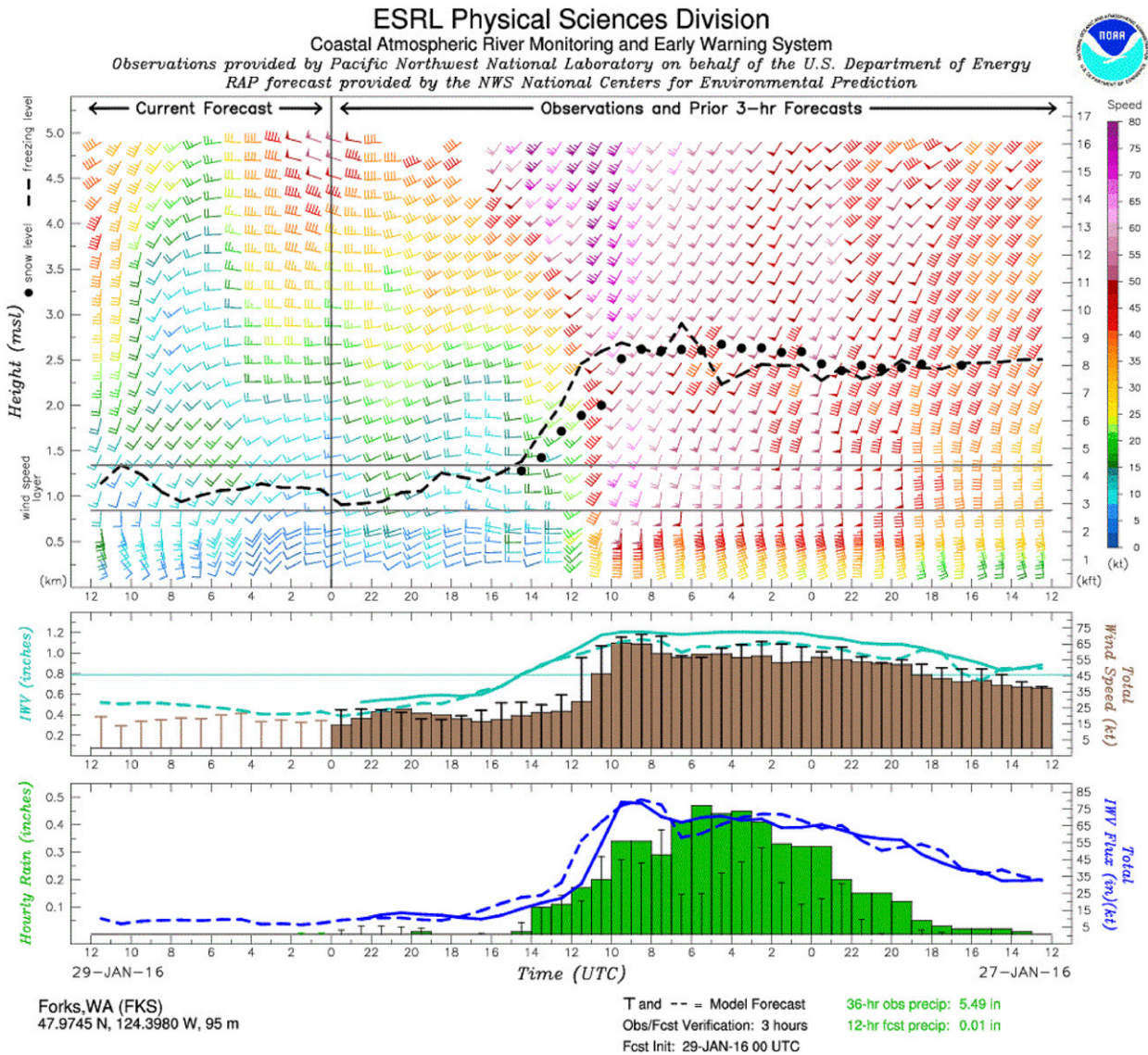


FIG. 1. Water vapor flux tool for the FKS station from 1200 UTC 27 Jan 2016 to 1200 UTC 29 Jan 2016 (note that time increases from right to left along the x axis). All values are hourly averages; all forecasts are 3-h forecasts. (top) Observed and forecasted wind profiles (the vertical bar in the top plot separates observed from forecasted variables; this also applies to the middle and bottom panels), observed snow level (black dots), and adjusted forecasted freezing level (black dashed line). (middle) Observed and forecasted winds averaged over the controlling wind layer, as designated by the horizontal lines in the top panel at 750 and 1250 m MSL. Observed and forecasted total wind speed is indicated by the brown bars and “T” bars, respectively. The middle panel also includes the observed and forecasted IWV (solid and dashed blue lines, respectively). (bottom) Observed and forecasted precipitation (green bars and “T” bars, respectively) and observed and forecasted IWV flux (solid and dashed dark blue lines, respectively). Flags = 50 kts (1 kt ≈ 0.51 m s<sup>-1</sup>); barbs = 10 kts; half-barbs = 5 kts; wind speed is color coded; the units for the rest of the measurements are indicated in the y-axis labels. While the tool is posted online in English units, the accompanying online text files and the analysis in this paper are in SI units.

Zhu et al. (2013), Bytheway and Kummerow (2015), and Gowan et al. (2018). Except for Gowan et al. (2018), these evaluations focused on precipitation east of the Rocky Mountains. Thus, our study is relatively unique in that it focuses on HRRR and RAP QPF evaluations in the complex terrain of the western United States.

Given that the operational versions of HRRR and RAP have been upgraded since this analysis was finalized, and the model developers continually work to improve the models, this study should be considered a snapshot of WVFT performance for one winter, providing insight and quantifications of errors in the forecasts presented in the WVFT.

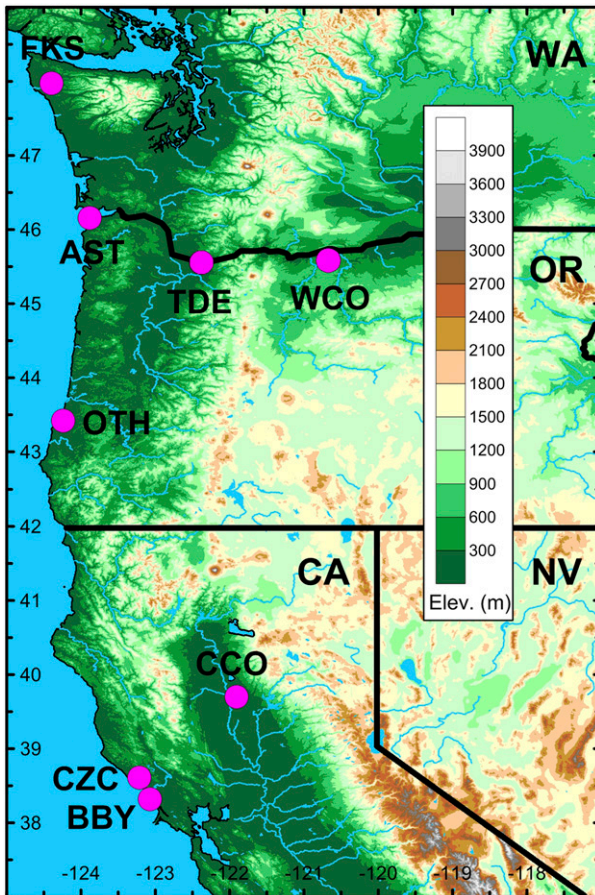


FIG. 2. Elevation (m MSL) map showing the location of the eight stations (pink dots) that supported the water vapor flux tool during winter 2015/16. Forks, WA (FKS), Astoria, OR (AST), Northbend, OR (OTH), Cazadero, CA (CZC), Bodega Bay, CA (BBY), Troutdale, OR (TDE), Wasco, OR (WCO), and Chico, CA (CCO). More information about the stations can be found in Table 1.

2. The water vapor flux tool

a. Details of the WVFT

The example of the NOAA WVFT in Fig. 1 shows a rainy period at FKS, a station near the Olympic Mountains of northwest Washington at an elevation of 95 m (Table 1, Fig. 2). The hourly wind profiles in the top panel of the tool and to the right of the vertical line were measured by a 449-MHz wind-profiling radar (profiler), and for this example, the predicted wind profiles to the left of the line were from the RAP operational runs. Two horizontal lines at 750 and 1250 m bracket the level of winds found most likely to drive orographic precipitation at CZC (Neiman et al. 2009; White et al. 2012, 2013). In their paper, Neiman et al. (2009) determined that winds between 140° and 320° azimuth and 850–1150 m MSL were key to producing extreme ( $\geq 10 \text{ mm h}^{-1}$ ) orographic precipitation

TABLE 1. For each station: name and 3-letter identifier, latitude, longitude, elevation, total precipitation for the study period, instrumentation relevant to this study, and supporting agency during the study period.

Station	Lat (°)	Lon (°)	Elev (m ASL)	Precipitation		Precipitation measurements	Profiler wind measurements	GPS/integrated water vapor measurements	Supporting agency
				total (mm)	1 Dec 2015–31 Mar 2016				
Forks, WA (FKS)	47.97	-124.40	95	1671.59		Tipping-bucket TR525USW	449 MHz	GpsTrimble NetR8	DOE
Astoria, OR (AST)	46.16	-123.88	3	1424.71		Tipping-bucket TR525USW	449 MHz	GpsTrimble NetR8	DOE
Cazadero, CA (CZC)	38.61	-123.22	478	1269.73		Tipping-bucket TE525	BBY winds	BBY IWV	HMT-West
Northbend, OR (OTH)	43.42	-124.24	5	1212.1		Tipping-bucket TR525USW	449 MHz	GpsTrimble NetRS at Coos Bay (CSB)	DOE
Troutdale, OR (TDE)	45.55	-122.39	12	743.18		Tipping-bucket heated TE525	915 MHz	GpsTrimble NetRS	DOE
Bodega Bay, CA (BBY)	38.32	-123.07	15	520.67		Tipping-bucket TE525	449 MHz	GpsTrimble NetR8	HMT-West
Chico, CA (CCO)	39.70	-121.91	42	405.88		Tipping-bucket TE525	915 MHz	GpsTrimble NetR7 until 21 Dec 2015; replaced with GpsTrimble NetRS on 18 Mar 2016	HMT-West
Wasco, OR (WCO)	45.59	-120.67	462	115.84		Tipping-bucket heated TE525	915 MHz	GpsTrimble NetRS at Wasco Airport (WCA)	DOE and HMT-West

at CZC. Neiman et al. (2009) referred to this layer as the *upslope wind layer*.

This upslope wind layer and its associated localized orographic precipitation enhancement is likely different at each site in the network. However, the upslope wind layer heights from BBY/CZC were used as a default at all other stations because an extended optimization analysis has yet to be conducted at these stations. The term *total wind layer* has been adopted for use in the WVFT at the other stations where the upslope component that drives the orographic precipitation has not yet been determined, and the total wind is used instead in the WVFT. To avoid these differences in definition, for the remainder of this paper we will refer to the total wind layer and the upslope wind layer both as the *controlling wind layer*.

The upper panel also shows the observed snow level (dots) and forecasted freezing level (dashed line). The forecasted freezing level on the tool has been adjusted downward by 233 m, corresponding to the average offset between the measured snow level and the freezing level, based on White et al. (2010). The middle plot shows observed and forecasted wind speed (solid brown bars and “T” bars, respectively) in the controlling wind layer bracketed by the two horizontal lines in the upper panel. The middle panel also shows observed and forecasted IWV (solid and dashed lines, respectively) and the IWV threshold for possible extreme precipitation (blue horizontal line). The bottom panel shows observed and forecasted IWV flux (solid and dashed lines, respectively) and the observed and forecasted precipitation (solid green bars and T bars, respectively).

Precipitation forecasts are the main topic of this paper, but to understand these forecasts and to provide a context for the precipitation forecast analysis, we also assess the controlling wind layer, IWV, and IWV flux predictions. Text files listing all of the observations and 3-h model forecasts shown in the tool are also available on the WVFT website. We used these text files for the analysis in this paper.

### b. The observations

Figure 2 is a terrain map showing the locations of the eight sites where model forecasts were evaluated. Figure 3 shows the total observed and forecasted precipitation at each site for the study period (1 December 2015–31 March 2016). Table 1 gives more information about the sites, including total observed precipitation for the study period and the relevant station instrumentation. The stations are arranged from most (top) to least (bottom) precipitation for the 4-month study period; the top four stations were coastal stations that received more than 1200 mm of precipitation over the study period. For context, the Stage IV precipitation measurements

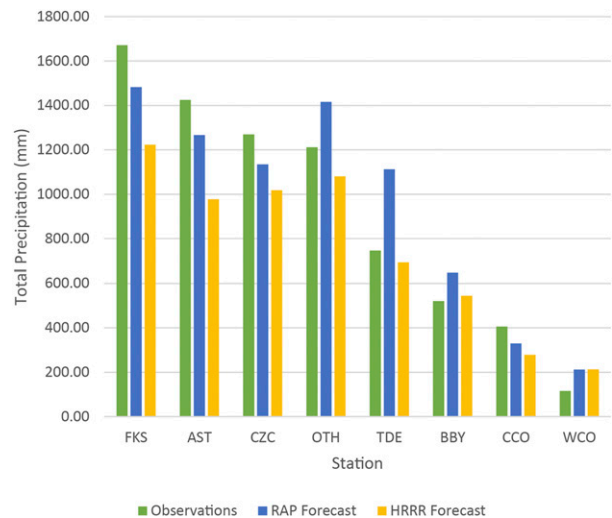


FIG. 3. Total observed and forecasted precipitation during the study period, 1 Dec 2015–31 Mar 2016 (mm). The observed precipitation comes from rain gauge data at each ARO. The forecasted precipitation is the sum of the HRRR and RAP operational 3-h forecasts at the grid point closest to each station.

(Lin and Mitchell 2005) for the study period are shown in Fig. 4. FKS and CZC were not only coastal stations, but also were elevated above sea level (95 and 478 m, respectively) and located on the windward side of the Coast Range. TDE is near the Columbia River, ~160 km from the coastline and located on the windward side of the Cascade Range. BBY, a coastal station, received less than half the precipitation received at nearby CZC because of lack of orographic lift relative to the CZC site, and it was south of the majority of the storm tracks (Fig. 4). Both CCO and WCO were inland stations. CCO was in the Central Valley of California. WCO was on an elevated plateau in Oregon and located in the lee of the Cascade Range, in a rain shadow.

Since the WVFT is a near-real-time tool, the observations and forecasts of key parameters were evaluated “as is,” with no additional quality control. However, 1-h observation gaps were filled using linear interpolation, mainly to account for the fact that the IWV observation was usually missing during the last hour of each day and we wanted to include as many IWV forecast–observation pairs as possible. Profiler derived winds were objectively quality controlled using an enhanced real-time version of the vertical–temporal continuity method described by Weber et al. (1993) and have a  $\sim 1 \text{ m s}^{-1}$  accuracy (Neiman et al. 2009). Precipitation gauge data are not quality controlled for this near-real-time tool. However, PSD uses Texas Electronics tipping-bucket rain gauges, without a wind screen, which have an accuracy of  $\pm 1.0\%$  up to  $50 \text{ mm h}^{-1}$  (<https://www.campbellsci.com/te525-1>), which far exceeds the maximum observed hourly

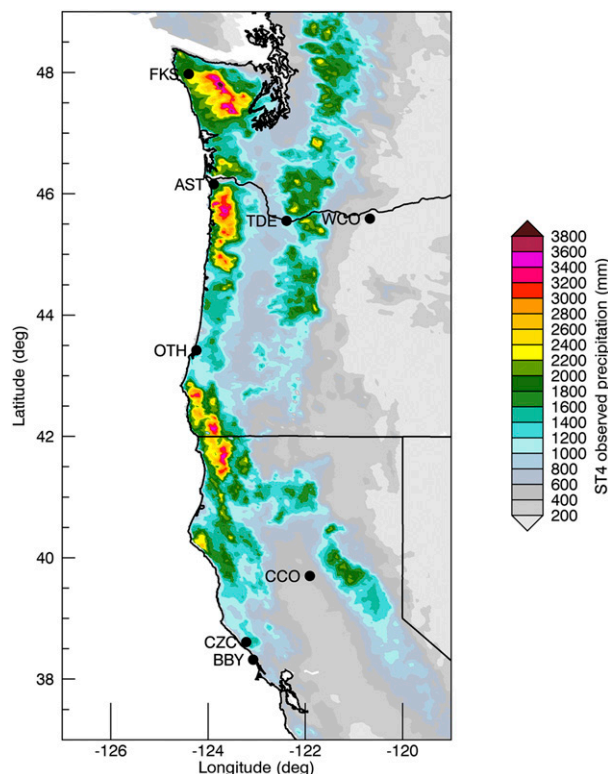


FIG. 4. Total Stage IV liquid equivalent precipitation analysis for the entire study period (1 Dec 2015–31 Mar 2016).

precipitation value of 19.3 mm at CZC that occurred during our study period. Without the wind screen, the gauges are likely undersampling the total precipitation, especially during windy conditions. However, at the station with the highest elevation and greatest rain rates, CZC, rain gauge data consistently compared well with a collocated disdrometer (Clark King 2018, personal communication). IWV measurements are derived from a GPS receiver collocated with the surface meteorology measurements (Bevis et al. 1992; Duan et al. 1996; Gutman et al. 2004; and White et al. 2013) and have a  $\sim 1$ -mm accuracy (Neiman et al. 2009). According to Gutman et al. (2004) GPS “. . . retrieval accuracies are comparable to, or better than, integrated rawinsonde measurements without the well-known problems at high humidity or low temperature.”

### c. The forecasts

The real-time display of the WVFT includes model output from NOAA/NCEP’s operational Rapid Refresh (RAP) assimilation and modeling system and the HRRR model, as well as experimental versions of one or both of these models (referred to as “rapexp” and “hrrrexp” on the tool web pages). The 3-h forecast values shown in the tool are from the model grid point

closest to the station. Since the experimental simulations were not always complete (i.e., there could be several missing hourly forecasts on any given day), we focused here on the versions of RAP and HRRR that were operational during the study period. All versions of the RAP [and its predecessor the Rapid Update Cycle (RUC)] are described in detail in Benjamin et al. (2016b), where the RAP and HRRR versions evaluated in this paper are referred to as RAP v2 and HRRR. The RAP v3 and HRRR v2 discussed in Benjamin et al. (2016b) became operational in August 2016.

The RAP and HRRR were updated hourly, with RAP producing hourly forecasts out to 18 h, and the HRRR producing forecasts out to 15 h, lending themselves well for a decision support tool such as the WVFT. The NOAA Gridpoint Statistical Interpolation (GSI) analysis system is used to assimilate radar reflectivity, and near-real-time data from rawinsondes, wind profilers (including the wind profiler data obtained at the PSD AROs), aircraft, surface meteorological observations, Geostationary Operational Environmental Satellite (GOES), and Global Positioning System (GPS) water vapor into the RAP (see Benjamin et al. 2016b, their Table 4, for details on the assimilated observations). The HRRR is nested within the RAP, with horizontal grid spacing of 3 km (vs. the RAP 13-km grid spacing). The RAP domain covered North America, and the HRRR domain covered the contiguous United States.

The HRRR is a convection-allowing model and forecasts also include the assimilation of radar reflectivity. According to Benjamin et al. (2016b), it is the 3D assimilation of radar reflectivity data, and the assimilation of 3D latent heating using 3D radar reflectivity data, or proxy reflectivity from lightning data, that sets the RAP and HRRR apart from other models. However, the assimilation of operational weather radar data in the west will be limited by radar beam blockage and other issues related to complex terrain (Bytheway et al. 2019 and Matrosov et al. 2014), as well as lack of upstream data over the Pacific Ocean. The development and evolution of the HRRR is also documented in Benjamin et al. (2016b).

## 3. Forecast evaluation metrics

### a. Issues in evaluating precipitation forecasts

Evaluations of precipitation forecasts can be problematic, starting with the representativeness and accuracy of the measurements used to evaluate the model. Colle et al. (1999, 2000, and references therein) explain many of the issues associated with precipitation measurements, including, but not limited to, excessive evaporation from heated rain gauges, wind effects, and freezing

TABLE 2. The number of observation–forecast pairs for each variable at each site for both models (HRRR/RAP).

	FKS	AST	CZC	OTH	TDE	BBY	CCO	WCO
Precipitation	2928/2928	2928/2928	2928/2928	2928/2928	2928/2928	2928/2928	2928/2928	2928/2928
IWV	2465/2469	2388/2391	2484/2489	2684/2684	2013/2019	2484/2489	707/707	2423/2425
Winds	2695/2689	2794/2793	2785/2791	2805/2798	2820/2822	2785/2791	2764/2771	2793/2795
IWV flux	2349/2348	2369/2372	2472/2476	2668/2662	2013/2019	2472/2476	679/679	2395/2397

precipitation. All of these problems can lead to measuring less precipitation than actually fell. Precipitation measurement accuracy issues are further complicated when looking at precipitation forecasts in complex terrain (Colle et al. 1999, 2000).

Yet even more complications arise when evaluating high-resolution model output (Mass et al. 2002; Grams et al. 2006), such as slight misplacements in forecast precipitation location that contribute to large errors when using traditional verification metrics (Ebert 2008; Han and Szunyogh 2016; Ikeda et al. 2013 and references therein) or the location of the precipitation is accurate but the timing is off (Ikeda et al. 2013). These types of situations are most likely associated with convective precipitation forecasts.

Despite these issues, we employ a point-to-point comparison of the observed and forecasted meteorological variables presented in the WVFT to ascertain how well the RAP and HRRR performed at each station in near-real time, because this method is compatible with the design of the WVFT. Our study period covers the months of December, January, February, and March, a period more likely to include widespread, day-long or multi-day-long events versus convective hit-or-miss-type events. Therefore, the precipitation events tended to be large-scale synoptically driven events, not mesoscale convective events. However, there are still frequently occurring mesoscale responses and accompanying precipitation enhancements in frontogenetic flow within these synoptic systems such as flow-parallel banded structures involving slant-wise convection that will impact forecasting skill at a single grid point.

### b. Metrics for evaluation

#### 1) RMSE, UNBIASED RMSE, AND STANDARD DEVIATIONS

The root-mean-squared error (RMSE), unbiased root-mean-squared error (RMSE<sub>UB</sub>), and standard deviation of the observations ( $\sigma_{\text{obs}}$ ) and forecasts ( $\sigma_f$ ) were computed for each hourly wind, precipitation, IWV, and IWV flux forecast–observation pair found in the study period, including hours without precipitation, at the eight stations, after Keyser and Anthes (1977) and Pielke (2002, p. 464). The RMSE gives an overall snapshot of forecast

performance by estimating an average error (Willmott 1982; Wilks 2006), however, RMSEs are known to be sensitive to large errors (Colle et al. 1999; Wilks 2006). The RMSE<sub>UB</sub> is computed the same way as the RMSE, after the means of the forecasts and the observations are removed from the forecasts and observations, respectively. Table 2 lists the number of forecast–observation pairs used for each variable at each station for both the HRRR and the RAP.

Keyser and Anthes (1977) suggest comparing the RMSE, RMSE<sub>UB</sub>, and  $\sigma_f$  to  $\sigma_{\text{obs}}$ . To demonstrate skill,  $\sigma_f$  and  $\sigma_{\text{obs}}$  should be approximately equal, and the RMSE and RMSE<sub>UB</sub> of the modeled variable should be less than  $\sigma_{\text{obs}}$ . Meeting these requirements implies that the magnitude of the errors in the forecasts and the spread of the modeled variables were reasonable compared to the observations.

In this analysis,  $\sigma_f$  and  $\sigma_{\text{obs}}$  were considered approximately equal if  $\sigma_f$  was within  $\pm 10\%$  of  $\sigma_{\text{obs}}$ . The range  $\pm 10\%$  was arbitrarily chosen by the authors. Throughout this paper the comparisons among the forecasted and observed standard deviations, RMSEs, and unbiased RMSEs will be referred to as the “K&A criteria,” after Keyser and Anthes (1977). Our results showed that in most cases there was very little difference between the RMSE and the RMSE<sub>UB</sub> values.

#### 2) MEAN ERROR

The mean error (ME) or the average difference between the hourly observations and forecasts for precipitation, IWV, IWV flux, and the  $u$  and  $v$  components of the wind were calculated for all observation and forecast pairs, including hours without precipitation. The mean errors indicate if the models tended to overpredict or underpredict the variables of interest, showing the direction (positive or negative) and magnitude of forecast biases over the study period (Wilks 2006). In a perfect forecast ME = 0.

#### 3) FREQUENCY BIAS SCORES

To investigate if the models did a better job of predicting some ranges of precipitation, IWV, winds, or IWV flux, relative to others, frequency bias scores [FBS; Eq. (1)] were computed for each variable at each station, using the hourly observation and model pairs, including

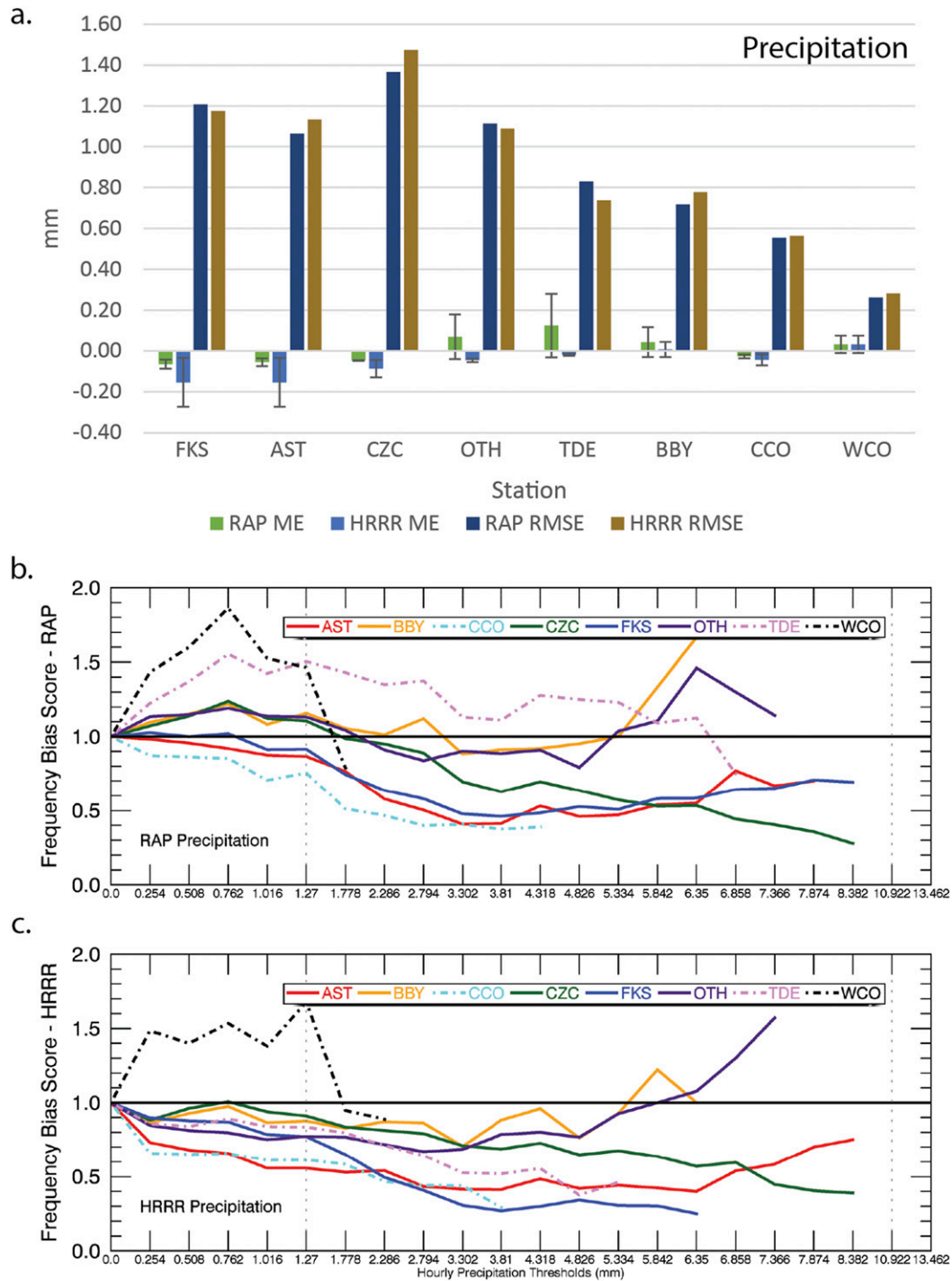


FIG. 5. The 3-h precipitation forecasts for each station: (a) mean error (ME) and root mean squared error (RMSE) for RAP and HRRR precipitation forecasts; (b) frequency bias scores for RAP precipitation forecasts; and (c) frequency bias scores for HRRR precipitation forecasts. The dash-dot lines represent inland stations. For reference, a solid black line at FBS = 1 is plotted; for FBS scores > 1 the model is overpredicting and for FBS scores < 1 the model is underpredicting. To compute the frequency bias scores a minimum of five matching observation-forecast pairs for each threshold value was required. The dashed vertical line at  $x = 1.27$  and  $10.933$  mm indicates a shift in the size of the increment between thresholds. Error bars at the 95% confidence interval have been added to the ME. FBS plots with 95% confidence intervals are in the [appendix](#).



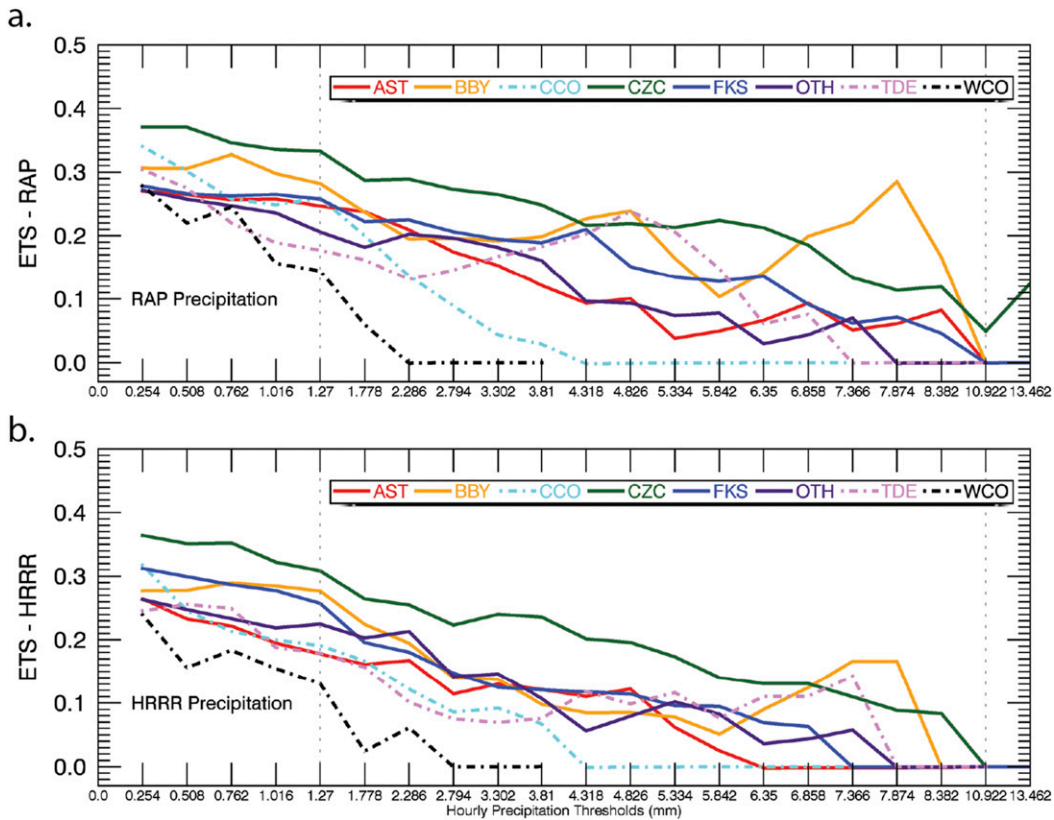


FIG. 6. Equitable threat scores for (a) RAP and (b) HRRR 3-h precipitation forecasts, using the same thresholds as the frequency bias scores. ETS plots with 95% confidence intervals are in the [appendix](#).

hours without precipitation. The thresholds for FBS calculations for each variable are indicated on the x axes of FBS plots (found in Figs. 5 and 7–10). A 95% confidence interval for the precipitation FBS were calculated using the bootstrapping method described by Hamill (1999). To account for serial dependence in the precipitation forecasts (as determined by the Spearman rank correlation), the bootstrapping was accomplished by randomly reshuffling the hourly forecasts in 24-h blocks. Figures showing the differences in FBS between the RAP and HRRR, with the 95% confidence interval, appear in the [appendix](#).

A score of FBS = 1 means there was no overprediction or underprediction. However, as Mass et al. (2002) point out, because of potential undercatchment issues in precipitation measurements in complex terrain, a slightly positive FBS for precipitation does not necessarily indicate an overprediction by the model. According to Colle et al. (1999) the FBS is not as sensitive to infrequent but large errors in forecasts or observations as the RMSE is, since all forecast–observation pairs are equally weighted in the FBS calculations while in calculating the RMSE, errors are squared before averaging, allowing large errors to dominate the result:

$$\text{bias score} = \frac{\# \text{ of forecasts} \geq \text{threshold}}{\# \text{ of observations} \geq \text{threshold}} \quad (1)$$

#### 4) EQUITABLE THREAT SCORE

The equitable threat score (ETS) is a skill score based on a 2 × 2 contingency table of yes/no observations and forecasts of a given event, in this case precipitation (Wilks 2006). The table accounts for four possible combinations of events: Precipitation was 1) forecast and observed; 2) forecast and not observed; 3) not forecast and observed; and 4) not forecast and not observed. The ETS is a measure of correct forecast *and* observed pairs relative to the number of times the event was forecast *or* observed, corrected for the number of correct “yes” forecasts that may occur by chance. The 95% confidence interval for the ETS was calculated using the bootstrapping method described in the previous section for the same hourly precipitation thresholds used in the frequency bias plots, as indicated by the values on the x axis of Fig. 6. Figures showing the differences in ETS between the RAP and HRRR, with a 95% confidence interval, appear in the [appendix](#). In a perfect forecast ETS = 1. A forecast with no skill will have an ETS ≤ 0.

### 5) PERCENTAGE OF OBSERVED PRECIPITATION

The percent of observed precipitation forecasted was evaluated.

Each metric described here only tells a part of the forecast evaluation story on its own. For instance, RMSE only gives information about the magnitude of the forecast error, the ME gives an overall bias with no indication of skill, the FBS only gives information about systematic overprediction or underprediction occurrences at specified thresholds, and the ETS gives information about forecast skill, but no information about forecast precipitation amounts (Wilks 2006). There is no single evaluation metric that can convey a model's forecast skill and accuracy, so it is important to include several evaluation metrics to assess a forecast. It is up to the forecaster to decide what metrics most directly apply to any given meteorological situation.

## 4. Results

### a. Precipitation forecasts

For the following discussion we break up the ARO stations into three groups: 1) stations where both the RAP and the HRRR underpredicted the study period total precipitation (FKS, AST, CZC, and CCO); 2) stations where the RAP overpredicted the study period precipitation and the HRRR underpredicted the precipitation (OTH and TDE); and 3) stations where both models overpredicted the study period precipitation (BBY and WCO) (Fig. 3).

#### 1) BOTH MODELS UNDERPREDICTED: FKS, AST, CZC, AND CCO

Both models underpredicted the total study period precipitation at the three stations that received the most precipitation over the study period (FKS, AST, and CZC) and at CCO, the inland Central Valley station (Fig. 3). The largest *percentage* underpredictions among all of the stations, not just in this group, occurred at AST and CCO by the HRRR, where the total HRRR forecast precipitation was 69% of the observed. The underpredictions at FKS, AST and CZC are consistent with the Lin et al. (2013) MM5 precipitation simulations in the Pacific Northwest (their Figs. 2 and 5), which indicated underpredictions at several coastal observation sites, including AST.

In Fig. 5a we see that the largest magnitude mean errors among all stations were the underpredictions at FKS and AST in the HRRR forecasts (0.153 mm in both cases). The largest RMSEs among all of the stations occurred at CZC (1.37 mm in the RAP, 1.47 mm in the HRRR).

TABLE 3. For the precipitation forecast from each model, the standard deviation of the forecast ( $\sigma_f$ ) is compared to the standard deviation of the observations ( $\sigma_{\text{obs}}$ ). The root-mean-squared error (RMSE) and the unbiased root-mean-squared error (RMSE<sub>UB</sub>) of the forecast are also compared to  $\sigma_{\text{obs}}$ . If  $\sigma_f \approx \sigma_{\text{obs}}$ , and if the RMSE and RMSE<sub>UB</sub> are  $< \sigma_{\text{obs}}$ , then the forecast at the station meets the Keyser and Anthes (1977) criteria, as explained in the text. Bolded text indicates the criteria were met.

		$\sigma_f/\sigma_{\text{obs}}$	RMSE/ $\sigma_{\text{obs}}$	RMSE <sub>UB</sub> / $\sigma_{\text{obs}}$
HRRR	FKS	0.7	0.9	0.9
	AST	0.8	1.0	1.0
	CZC	0.7	0.9	0.9
	OTH	1.0	1.0	1.0
	TDE	0.8	1.0	1.0
	BBY	1.1	1.1	1.1
	CCO	0.7	1.0	1.0
RAP	WCO	1.2	1.2	1.2
	FKS	<b>0.9</b>	<b>0.9</b>	<b>0.9</b>
	AST	0.8	0.9	0.9
	CZC	0.8	0.8	0.8
	OTH	1.0	1.1	1.0
	TDE	1.1	1.1	1.1
	BBY	1.1	1.0	1.0
CCO	0.8	0.9	0.9	
	WCO	1.1	1.1	1.1

RAP precipitation forecasts at FKS had a standard deviation, RMSE, and RMSE<sub>UB</sub> of the right magnitudes to meet the K&A criteria (Table 3), implying that the spread of the forecast values and magnitude of errors were reasonable compared to the observations. However, even though the RMSE and RMSE<sub>UB</sub> were  $< \sigma_{\text{obs}}$ , they were approaching the magnitude of  $\sigma_{\text{obs}}$ . For precipitation this was the *only* site among *all* stations where both models met the K&A criteria. The standard deviations at the other three sites, AST, CZC, and CCO, were too small to be considered approximately equal to those of the measurements, indicating that the models were not properly representing the variability, or spread, among the hourly precipitation forecasts at these stations.

The FBS (Figs. 5b,c and the appendix) give an indication of how each model did at predicting hourly precipitation totals at each site, binned by threshold values. For the small precipitation rates the RAP (Fig. 5b) tended to overpredict precipitation at CZC and underpredict precipitation at AST and CCO. The RAP FBS for the low values at FKS were close to 1, a possible explanation for the RAP FKS precipitation forecast meeting the K&A criteria—the model did well at predicting the smaller precipitation amounts at FKS. As hourly observed precipitation amounts increased, the RAP forecasts underpredicted precipitation at FKS, AST, CZC, and CCO; in fact, at the greatest observed precipitation values these stations had the lowest FBS among all stations.

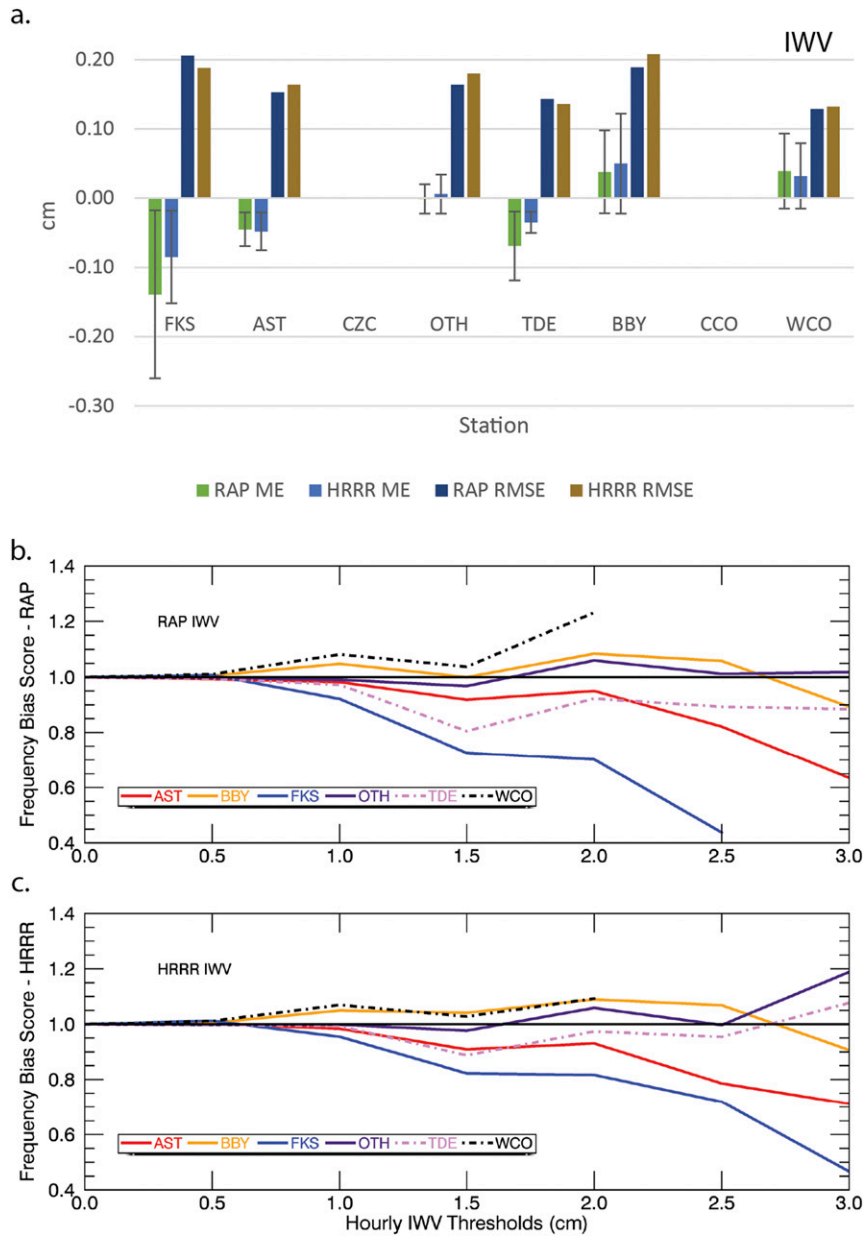


FIG. 7. As in Fig. 5, but for integrated water vapor.

The HRRR FBS (Fig. 5c) showed that precipitation at these four stations tended to be underpredicted at the greatest observed values, leading to the greatest underpredictions of the study period, as noted above. In fact, FBS for both models reached <0.5 for some observed values, meaning that more than half of the time these hourly values were underforecast. The ETS for these stations showed a decline in skill as the observed precipitation increased (Fig. 6). At these stations the differences in FBS and ETS between the two models were not statistically significant at most threshold values

as indicated by the 95% confidence interval plots shown in the appendix.

2) RAP OVERPREDICTED, HRRR UNDERPREDICTED: OTH AND TDE

The RAP overpredicted the study period precipitation at TDE and OTH by 149% and 117%, respectively. The HRRR study-period forecasts for TDE and OTH, however, were close to the observed precipitation values (93% and 89%, respectively). At both stations the MEs were positive for the RAP and negative for the HRRR

(Fig. 5a). K&A criteria analysis showed that the RMSEs and  $RMSE_{UBS}$  were too large for both models at both locations, and the HRRR  $\sigma_f$  at TDE was too small (Table 3).

RAP FBS for TDE (Fig. 5b and the appendix) remained  $>1$  at almost all values, consistent with the larger positive ME, the overall overprediction of precipitation at TDE for the study period, and, as will be shown later,  $FBS > 1$  for IWV flux. RAP frequency bias scores for OTH indicated overpredictions of precipitation at the smallest and largest hourly amounts, and a bias score  $< 1$  for hourly values in the middle range. HRRR FBS for TDE showed underpredicted precipitation at all observed values, and underpredicted precipitation at OTH except for the largest observed values. The ETS generally showed a decline in skill as precipitation increased at these stations, but the decline at TDE was not steady. The differences in the RAP and HRRR FBS at TDE and OTH were statistically significant at the smaller precipitation thresholds; the differences in ETS scores were not statistically significant (see the appendix).

For the hours of heaviest precipitation at TDE, the RAP was likely to overpredict precipitation, whereas the HRRR was more likely to underpredict it. For OTH, the results were mixed for values near 1.93 mm (close to 1 in the RAP forecasts and  $<1$  for the HRRR forecasts), fitting in with the MEs and study period totals.

### 3) BOTH OVERPREDICTED: BBY AND WCO

BBY and WCO were the only stations where both the RAP and the HRRR overpredicted the study period precipitation total (Fig. 3) and had positive MEs (Fig. 5a). The largest percentage overpredictions of precipitation among all of the stations occurred at WCO, with both models predicting 184% of the observed precipitation. However, the HRRR study period total at BBY was close to observed (105%). Looking at the FBS plots, the RAP and HRRR overprediction of precipitation at WCO was very apparent (Figs. 5b,c and the appendix). The RMSEs and  $RMSE_{UBS}$  for the BBY and WCO forecasts were too large to meet the K&A criteria (Table 3). The differences in FBS and ETS between the two models were not statistically significant at these two sites (see the appendix).

### 4) SUMMARY OF PRECIPITATION FORECAST ANALYSIS

Our forecast evaluation metrics showed that neither the RAP nor the HRRR precipitation forecasts, with grid spacings of 13 and 3 km, respectively, had a strong performance advantage over the other in forecasting precipitation at our eight stations, at least using these traditional nearest-point metrics. ETS for both models showed an overall decline in forecast skill as the observed

TABLE 4. As in Table 3, except for integrated water vapor.

		$\sigma_f/\sigma_{obs}$	$RMSE/\sigma_{obs}$	$RMSE_{UB}/\sigma_{obs}$	
HRRR	FKS	<b>0.9</b>	<b>0.3</b>	<b>0.3</b>	
	AST	<b>1.0</b>	<b>0.3</b>	<b>0.3</b>	
	CZC	—	—	—	
	OTH	<b>1.0</b>	<b>0.3</b>	<b>0.3</b>	
	TDE	<b>1.0</b>	<b>0.3</b>	<b>0.3</b>	
	BBY	<b>1.0</b>	<b>0.3</b>	<b>0.3</b>	
	CCO	—	—	—	
	WCO	<b>1.0</b>	<b>0.3</b>	<b>0.3</b>	
	RAP	FKS	<b>0.9</b>	<b>0.4</b>	<b>0.3</b>
		AST	<b>1.0</b>	<b>0.3</b>	<b>0.2</b>
CZC		—	—	—	
OTH		<b>1.0</b>	<b>0.3</b>	<b>0.3</b>	
TDE		<b>1.0</b>	<b>0.3</b>	<b>0.2</b>	
BBY		<b>1.0</b>	<b>0.3</b>	<b>0.3</b>	
CCO		—	—	—	
WCO		<b>1.0</b>	<b>0.3</b>	<b>0.3</b>	

hourly precipitation amount increased, and FBS for precipitation echoed this declining skill.

The RAP tended to overpredict smaller hourly amounts, except for FKS, AST, and CCO, although, if considering potential undercatchment issues (Colle et al. 1999 and 2000), the slightly positive FBS may not necessarily indicate an overforecast (Mass et al. 2002, Gowan et al. 2018). For instance, if a rain gauge consistently did not properly collect the precipitation that fell at a location because of strong winds, an FBS slightly  $>1$  (e.g., as seen in the RAP forecasts for OTH, BBY, and CZC at the smallest observed amounts) may indicate accurate precipitation forecasts at these amounts. The HRRR tended to underpredict the average hourly amounts at all stations except WCO. FBS became worse at the larger hourly precipitation values for some stations, with both models underpredicting larger precipitation values at AST, CCO, CZC, and FKS. Accurately forecasting these large precipitation rates is critical for forecasting heavy precipitation events at these sites, which could potentially lead to flash flooding. WCO, the station receiving the least amount of precipitation, was an outlier in that both models tended to overpredict precipitation at most observed amounts.

Only RAP precipitation forecasts at FKS met the K&A criteria, but in all cases, including this one, the RMSEs were large compared to the  $\sigma_{obs}$ . Except for the HRRR forecast at WCO, the standard deviations of the forecasts that did not meet the K&A criteria tended to be too small, indicating a lack of spread in the precipitation forecasts, probably because of the underprediction of larger hourly values. Overall, the average errors in the precipitation forecasts were large and the spread among the values was small relative to  $\sigma_{obs}$ .

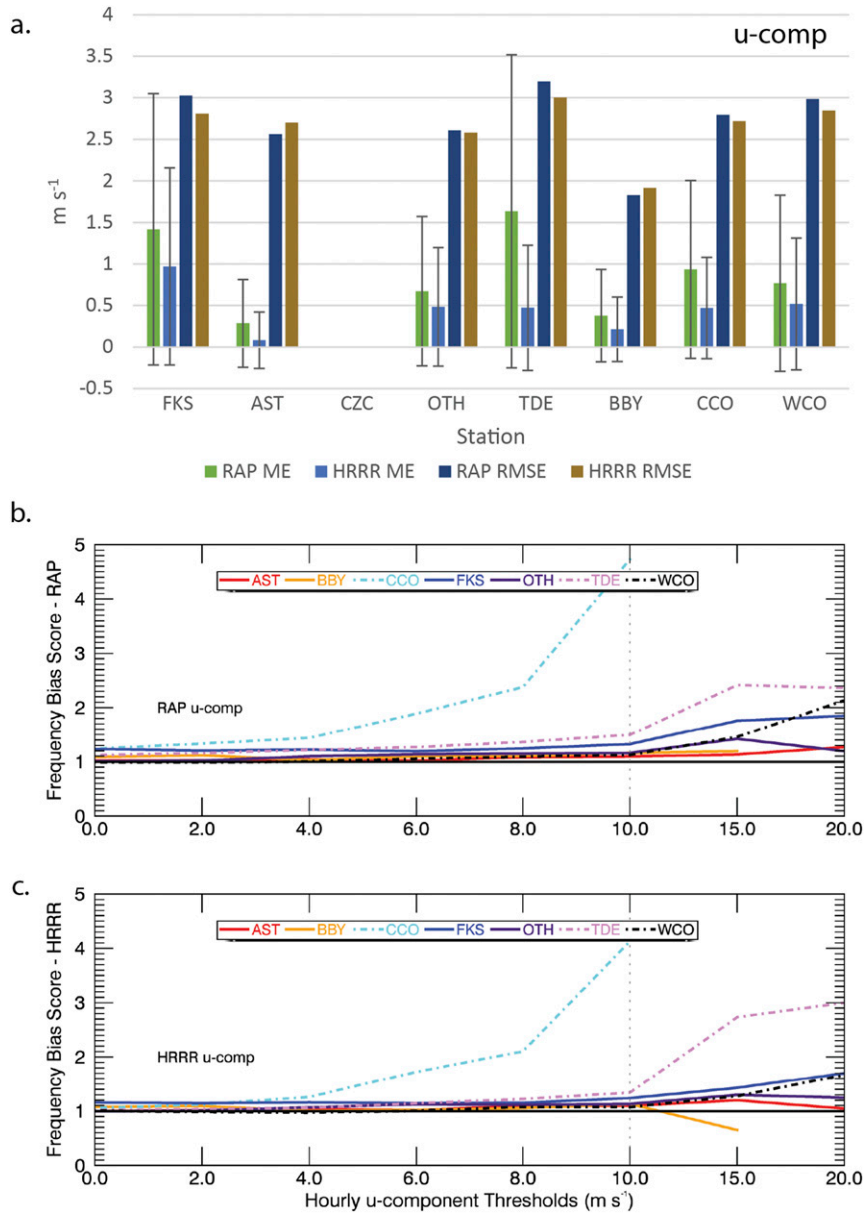


FIG. 8. As in Fig. 5, but for the  $u$  component of the wind.

Model-predicted precipitation is directly related to the water vapor flux to a site, which is the product of the vertically integrated water vapor and the controlling-layer wind speeds. In the following subsections we evaluate these components.

*b. First component: Integrated water vapor forecasts*

We first evaluate integrated water vapor forecasts, which indicate how well the models predicted the moisture available for precipitation production. Figure 7a shows the MEs and RMSEs for both RAP and HRRR forecasts of IWV at all stations that had at least 2000 h

(out of 2928 total hours) of IWV measurements (IWV analysis was not performed at two sites, CZC, because IWV was not measured there, and CCO because there were too many data outages during the study period). FKS, AST, and TDE IWV forecasts were negatively biased, whereas BBY and WCO forecasts were positively biased. BBY and WCO were the only stations where both the RAP and HRRR overforecasted the total precipitation.

FBS for IWV (Figs. 7b,c) indicated how hourly IWV predictions compared to hourly IWV measurements at a range of thresholded values. For both models, hourly

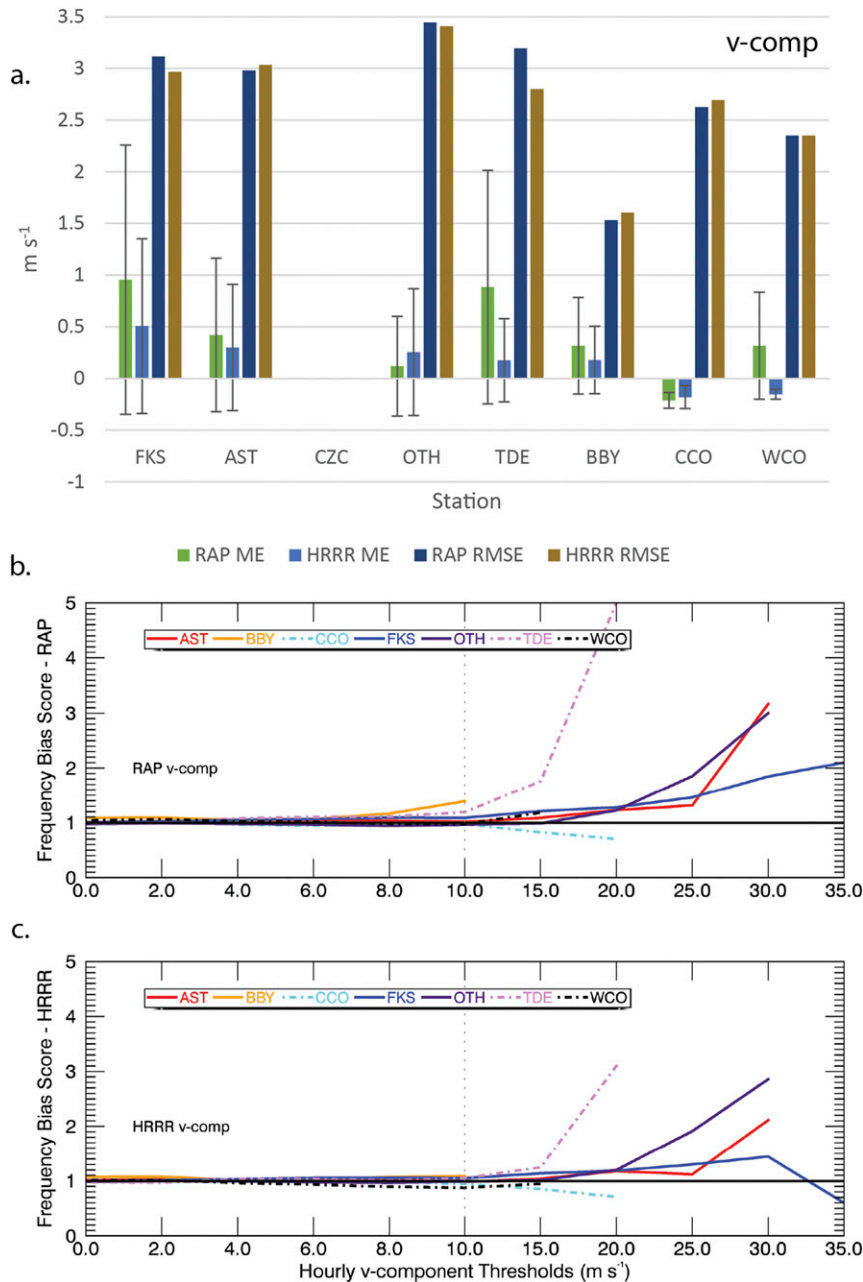


FIG. 9. As in Fig. 5, but for the  $v$  component of the wind.

averaged IWV forecasts of  $\leq 0.5$  cm were excellent at all sites, with  $FBS \approx 1$ . Model performance diverged at values  $> 0.5$  cm. FKS, AST, and TDE bias scores showed IWV to be underpredicted by both models at most values  $> 1.0$  cm, with FKS having the worst performance (the most underforecasting) for both models. OTH and BBY IWV FBS were close to 1 at most values observed. WCO was the only station showing an overprediction of IWV at all values  $> 0.5$  cm, which likely led to the overforecasting of precipitation at this station.

IWV forecasts by both models for all stations had RMSE,  $RMSE_{UB}$ , and  $\sigma_f$  values that compared well to the standard deviations of the observations, indicating that the forecasts were realistic (Table 4). Our IWV MEs were consistent with Wick et al. (2013) who found similar results when evaluating operational ensemble forecast systems. The greatest problems in the IWV forecasts were that neither the RAP nor the HRRR forecasted larger ( $> 2.0$  cm) IWV values at FKS and AST, the two stations that received the heaviest

TABLE 5. As in Table 3, except for the  $u$  component of the wind.

		$\sigma_f/\sigma_{\text{obs}}$	RMSE/ $\sigma_{\text{obs}}$	RMSE <sub>UB</sub> / $\sigma_{\text{obs}}$	
HRRR	FKS	<b>1.0</b>	<b>0.5</b>	<b>0.4</b>	
	AST	<b>1.1</b>	<b>0.4</b>	<b>0.4</b>	
	CZC	—	—	—	
	OTH	<b>1.1</b>	<b>0.4</b>	<b>0.4</b>	
	TDE	<b>1.1</b>	<b>0.4</b>	<b>0.4</b>	
	BBY	<b>1.0</b>	<b>0.4</b>	<b>0.4</b>	
	CCO	<b>1.1</b>	<b>0.8</b>	<b>0.8</b>	
	WCO	<b>1.1</b>	<b>0.4</b>	<b>0.4</b>	
	RAP	FKS	<b>1.0</b>	<b>0.5</b>	<b>0.4</b>
		AST	<b>1.0</b>	<b>0.4</b>	<b>0.4</b>
CZC		—	—	—	
OTH		<b>1.1</b>	<b>0.4</b>	<b>0.4</b>	
TDE		<b>1.0</b>	<b>0.5</b>	<b>0.4</b>	
BBY		<b>1.0</b>	<b>0.4</b>	<b>0.4</b>	
CCO		<b>1.1</b>	<b>0.8</b>	<b>0.8</b>	
WCO		<b>1.1</b>	<b>0.4</b>	<b>0.4</b>	

TABLE 6. As in Table 3, except for the  $v$  component of the wind.

		$\sigma_f/\sigma_{\text{obs}}$	RMSE/ $\sigma_{\text{obs}}$	RMSE <sub>UB</sub> / $\sigma_{\text{obs}}$	
HRRR	FKS	<b>1.1</b>	<b>0.3</b>	<b>0.3</b>	
	AST	<b>1.0</b>	<b>0.4</b>	<b>0.4</b>	
	CZC	—	—	—	
	OTH	<b>1.1</b>	<b>0.4</b>	<b>0.4</b>	
	TDE	1.2	0.5	0.5	
	BBY	<b>1.0</b>	<b>0.4</b>	<b>0.4</b>	
	CCO	<b>0.9</b>	<b>0.3</b>	<b>0.3</b>	
	WCO	<b>1.0</b>	<b>0.4</b>	<b>0.4</b>	
	RAP	FKS	<b>1.1</b>	<b>0.4</b>	<b>0.3</b>
		AST	<b>1.1</b>	<b>0.4</b>	<b>0.4</b>
CZC		—	—	—	
OTH		<b>1.1</b>	<b>0.4</b>	<b>0.4</b>	
TDE		1.3	0.6	0.6	
BBY		<b>1.0</b>	<b>0.4</b>	<b>0.4</b>	
CCO		<b>0.9</b>	<b>0.3</b>	<b>0.3</b>	
WCO		<b>1.0</b>	<b>0.4</b>	<b>0.4</b>	

precipitation, at a high enough frequency, as indicated by FBS.

### c. Second component: Wind forecasts

Next we look at the wind forecasts in the elevated controlling wind layer, the layer responsible for advecting moisture toward the complex terrain of the coast. For the analysis of HRRR and RAP wind forecasts, the hourly averaged winds in the controlling layer were decomposed into  $u$  and  $v$  components, and the ME, RMSE, and RMSE<sub>UB</sub> were calculated for both components in each model. Both models had positive mean errors in the  $u$  component of the wind at all stations (Fig. 8a), the RAP having larger mean errors than the HRRR in every case. For the  $v$ -component winds (Fig. 9a), all stations except CCO and WCO had positive mean errors in both models, with CCO the only station with negative mean errors in both models.

FBS were calculated for the positive values of the  $u$  and  $v$  components of the wind (Figs. 8b,c and 9b,c), appropriate for the southwesterly flows associated with high-precipitation events along the west coast. The HRRR and RAP FBS for the  $u$  and  $v$  components of the wind were  $\approx 1$  for values  $\leq 10 \text{ m s}^{-1}$ , with the exception of CCO, which saw large  $u$ -component overpredictions in both models as wind speed increased (Figs. 8b,c and 9b,c). The models performed less accurately at the greater wind speeds that are associated with greater precipitation events, tending to overpredict these greater values at most stations.

HRRR and RAP forecasts of the  $u$  component met the K&A criteria at all stations (Table 5), meaning that the errors and spread in the modeled winds at these stations, averaged over the controlling wind layer, were reasonable when compared to  $\sigma_{\text{obs}}$ , although at CCO,

the RMSE and RMSE<sub>UB</sub> were approaching  $\sigma_{\text{obs}}$ . For the  $v$  component of the wind (Table 6),  $\sigma_f$  for the HRRR and RAP TDE forecasts slightly exceeded  $\sigma_{\text{obs}}$ , making TDE the only station not meeting the K&A criteria for the wind forecasts. CCO  $v$ -component forecasts were better than the  $u$ -component forecasts.

CCO lies in the Central Valley of California (Fig. 2). It is possible that the models failed to characterize the depth of the Sierra barrier jet accurately (Neiman et al. 2014), leading to an overprediction of the  $u$  component of the wind and a slight underprediction of the  $v$  component. The low-level easterly gap flows that dominate at TDE may not be adequately forecast by the models because of complex interactions between this gap flow and the moist, westerly overrunning flow originating over the Pacific that frequently occurs with land-falling storms. The  $v$ -component winds at FKS, AST, and OTH tended to be overpredicted at larger values. Overall, the models had good FBS at wind component values  $\leq 10 \text{ m s}^{-1}$ , but tended to overforecast the winds  $> 10 \text{ m s}^{-1}$  at some stations.

### d. Integrated water vapor flux forecasts

Integrated water vapor flux is the product of wind speed in the controlling wind layer and IWV; thus the errors in the wind and IWV forecasts will have an impact on the IWV flux forecasts. The IWV flux MEs for both models were positive in most cases, with only FKS having very small negative MEs (Fig. 10). The 3-h forecasts from both models at FKS and AST underpredicted IWV (as shown by the IWV ME and FBS plots, Fig. 7), but overpredicted the winds at greater values (as seen in the  $u$ - and  $v$ -component ME and FBS plots, Figs. 8 and 9). This offsetting of forecast errors produced very small MEs in IWV flux predictions at FKS and AST. TDE also

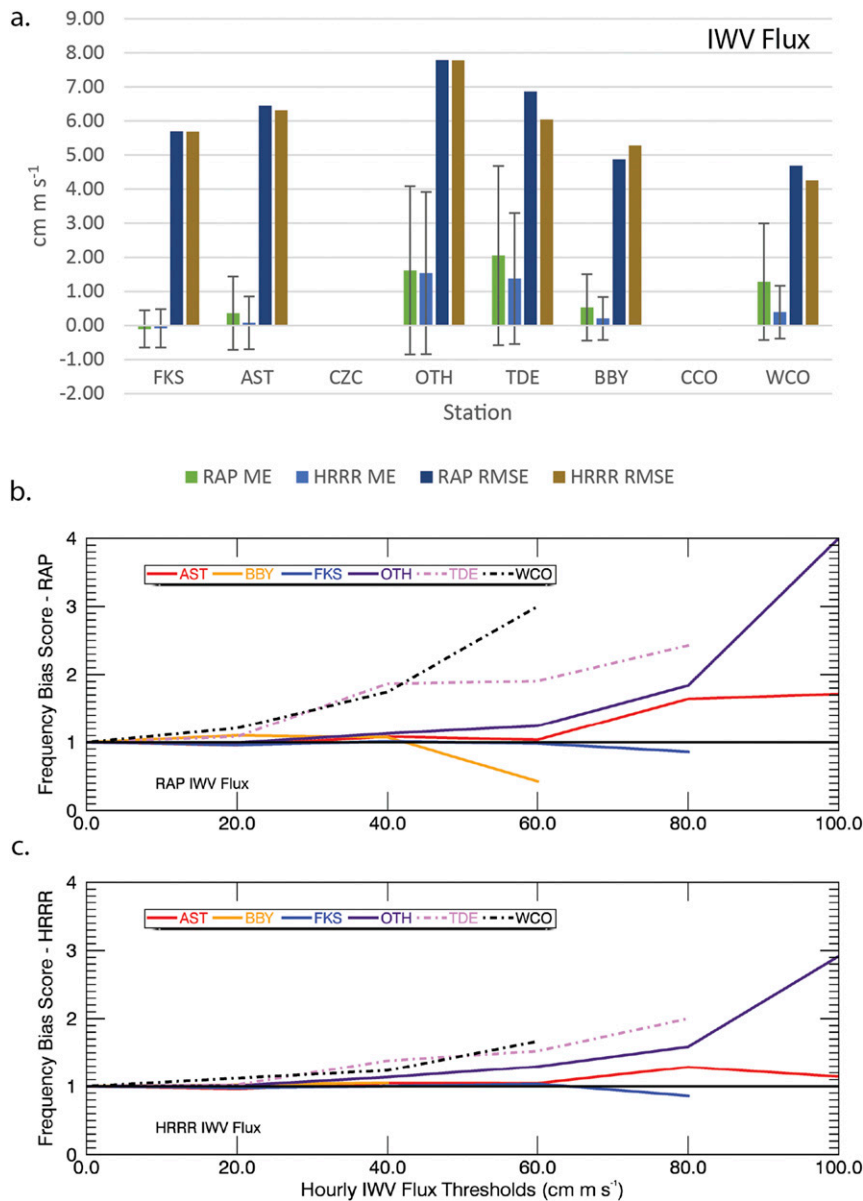


FIG. 10. As in Fig. 5, but for the integrated water vapor flux.

had underpredictions of IWV compensated for by the overprediction of winds.

At all stations the RMSE and RMSE<sub>UB</sub> values for both models and all stations were less than  $\sigma_{obs}$  (Table 7). The RAP IWV flux forecasts at OTH, TDE, and WCO did not meet the K&A criteria because  $\sigma_f > \sigma_{obs}$ , indicating that these IWV flux forecasts had more spread than the observations. The problems with the IWV flux forecasts at OTH, TDE, and WCO become more apparent when looking at the FBS plots of the IWV flux forecasts for both models (Figs. 10b,c). There was a rapid increase in RAP overpredictions of IWV flux at TDE and WCO as observed IWV flux values

increased. OTH also had larger overpredictions of the larger values of IWV flux. This overprediction of IWV flux possibly explains the RAP forecasting too much precipitation over the study period at OTH, TDE, and WCO. Only RAP IWV flux forecasts at BBY showed a tendency to be underpredicted at greater observed values. IWV flux forecasts  $\leq 20 \text{ cm m s}^{-1}$  had FBS  $\approx 1$  in both models at all stations. Therefore, with the exception of RAP IWV flux forecasts at BBY, this analysis indicates that the models advected sufficient moisture to the stations (in some instances too much moisture) by the controlling wind layer, providing an abundance of IWV flux forcing for the precipitation forecasts.



TABLE 7. As in Table 3, except for integrated water vapor flux.

		$\sigma_f/\sigma_{\text{obs}}$	RMSE/ $\sigma_{\text{obs}}$	RMSE <sub>UB</sub> / $\sigma_{\text{obs}}$
HRRR	FKS	<b>1.0</b>	<b>0.3</b>	<b>0.3</b>
	AST	<b>1.0</b>	<b>0.3</b>	<b>0.3</b>
	CZC	—	—	—
	OTH	<b>1.1</b>	<b>0.4</b>	<b>0.4</b>
	TDE	<b>1.1</b>	<b>0.4</b>	<b>0.4</b>
	BBY	<b>1.0</b>	<b>0.4</b>	<b>0.4</b>
	CCO	—	—	—
	WCO	<b>1.1</b>	<b>0.4</b>	<b>0.4</b>
RAP	FKS	<b>1.0</b>	<b>0.3</b>	<b>0.3</b>
	AST	<b>1.1</b>	<b>0.3</b>	<b>0.3</b>
	CZC	—	—	—
	OTH	1.2	0.4	0.4
	TDE	1.2	0.5	0.5
	BBY	<b>1.0</b>	<b>0.4</b>	<b>0.4</b>
	CCO	—	—	—
	WCO	1.2	0.5	0.5

## 5. Summary

In this paper we evaluated forecasted variables for an online near-real-time water vapor flux tool: IWV, controlling-layer winds, IWV flux, and precipitation. The tool is meant to provide situational awareness and decision support for detecting high-precipitation events in the western United States where orographic forcing is dominant. The goal of the paper was to characterize and quantify the errors in the tool forecasts, with a particular interest in the operational RAP and HRRR 3-h lead-time precipitation forecasts.

The key result is that the precipitation forecasts shown in the WVFT are likely to be underpredictions for larger hourly values of precipitation, despite IWV flux forecasts that approximately equaled or exceeded the observed in most cases, including at and above the critical  $25 \text{ cm m s}^{-1}$  IWV flux threshold for potential extreme precipitation. Precipitation forecasts only met the K&A criteria at one site, whereas the winds, IWV, and IWV flux met the criteria at most or all sites. Both FKS and AST (the stations with the most precipitation) did not have an adequate amount of IWV forecast (as defined by an FBS  $\geq 1$  and an ME  $\geq 0$ ) by either the RAP or the HRRR. In both cases the models did overpredict the winds, resulting in IWV flux forecasts with an FBS  $> 1$  or  $\approx 1$ , indicating adequate moisture flux was available in the model, relative to the observations. Yet the precipitation was underforecast by both models at most values at these stations, with the FBS and ETS decreasing as the observed precipitation values increased. This implied that the underforecast of IWV had a greater impact on the accuracy of the precipitation forecast than the resultant IWV flux forecast at these stations. There was a similar situation at TDE (IWV was underforecast

but the IWV flux was overforecast), but there were mixed results with the RAP overforecasting and the HRRR underforecasting the precipitation. At the opposite end of the spectrum, at WCO, the IWV and IWV flux were overpredicted by both models, and both models overpredicted the precipitation at this station. Mixed results occurred at OTH with the RAP overpredicting IWV flux and precipitation at some values.

According to Ralph et al. (2010), precipitation forecasts in complex terrain can be erroneous because of errors in forecasting lower-tropospheric wind direction or the structure of blocking phenomena (e.g., a barrier jet), and cloud microphysics. We have shown that in the case of the WVFT, the controlling-layer wind forecasts were reasonable at component speeds  $\leq 10 \text{ m s}^{-1}$ , as indicated by the K&A criteria and FBS, with the exception that the barrier jet in the Central Valley, affecting forecasts at CCO, and the gap flows at TDE may not be well represented in the models. Problems with the wind forecasts at greater speeds, as discussed above, could have contributed to overforecasts of the precipitation field at some stations.

Possible weaknesses in the models may include resolution of the terrain, the representation of the thermodynamics of approaching storm systems, and the vertical distribution of moisture. Improved data assimilation to more accurately capture moisture and wind profiles for model initialization and improved modeled physical processes to more accurately resolve features such as barrier jets could be considered. The lack of upstream measurements of wind and water vapor may also be an issue for precipitation forecasts in the western United States.

Potential errors in the vertical velocity field (which is an important input for microphysical schemes) will result if the low-level thermodynamic stability is inaccurately forecast by the models. That is, for a site that is dominated by orographic forcing (like CZC), the correctly forecasted upslope IWV flux will only result in the correct QPF if the vertical velocity response is accurately simulated. Likewise, the vertical profiles of vertical velocity and specific humidity can be inaccurately forecasted, even though the upslope IWV flux is correctly forecasted. In this case, the input for a microphysics scheme will result in an erroneous QPF. In addition, for sites that aren't dominated by orographic forcing, like AST or FKS, using the IWV flux results in even more ambiguities, since the forcing at these sites is not predominantly and explicitly provided by the IWV flux. Even though it is usually coincidental that other forcings (like frontal and synoptic-dynamic) are phased with the total IWV-flux plume, these forcings and the wind field (associated with the flux) are subject to both magnitude and phase lags and additional forecast errors.

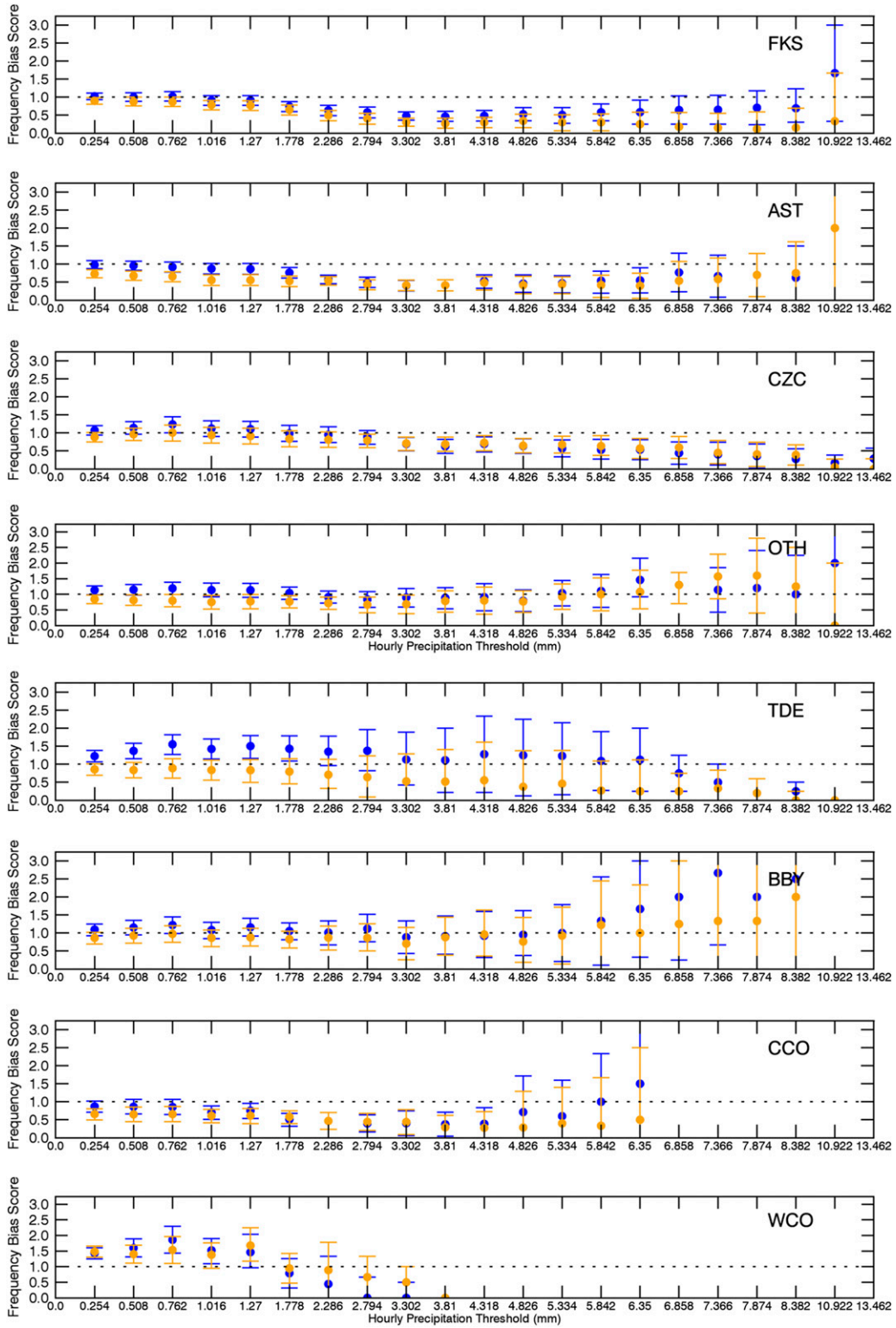


FIG. A1. Frequency bias plots with 95% confidence intervals for RAP (blue) and HRRR (orange) precipitation forecasts.

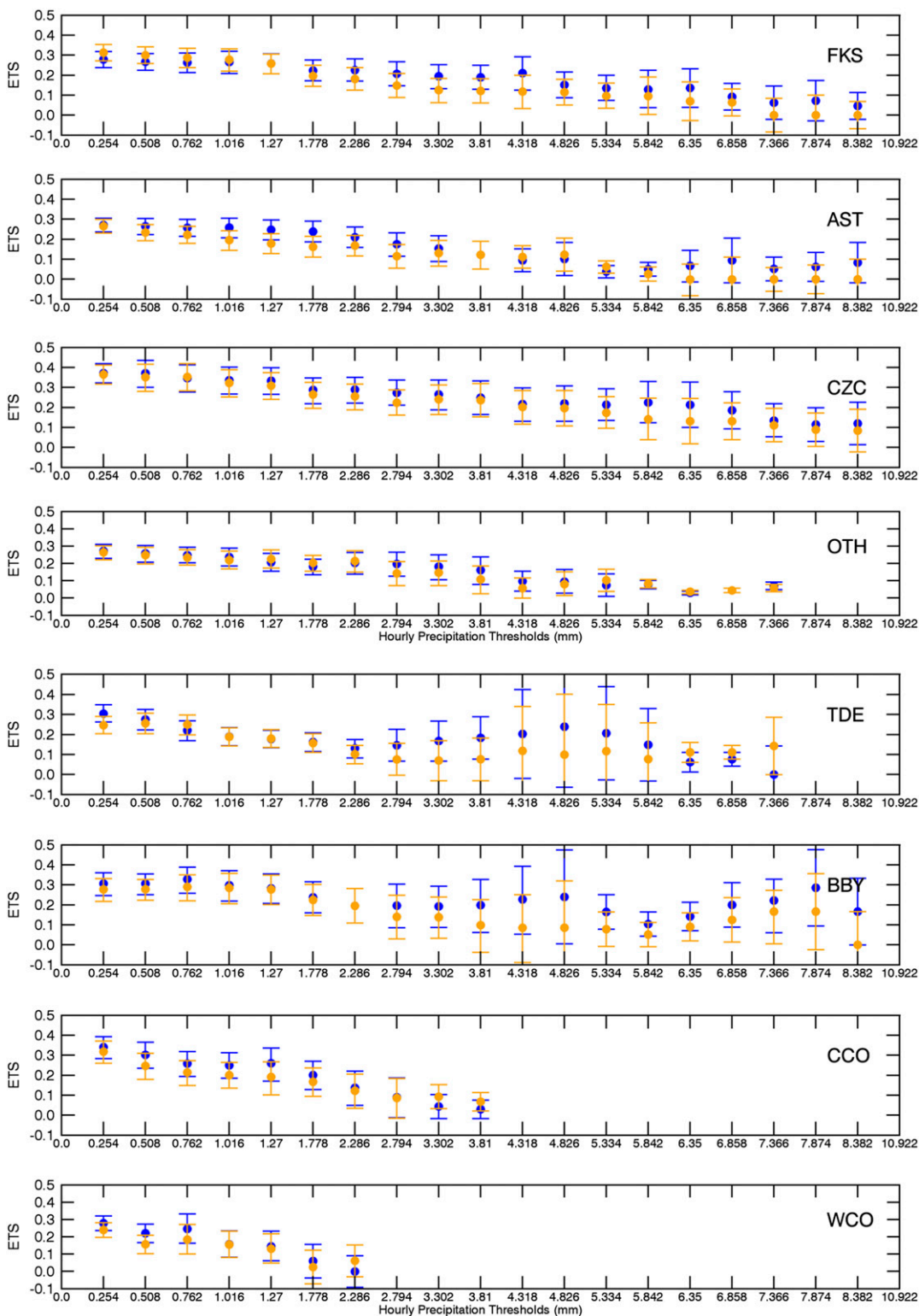


FIG. A2. Equitable threat score plots with 95% confidence intervals for RAP (blue) and HRRR (orange) precipitation forecasts.

We cannot address microphysics in this paper, but this analysis, and others, points to the need for additional research to take place regarding the cloud microphysics packages included in the RAP and HRRR. Previous studies have found that how a forecast model handles graupel aloft (Colle et al. 2005; Lin and Colle 2009), the intercept for snow-size distribution, and the number of CCN (Colle et al. 2005) can cause errors in the forecast of precipitation at the surface. Gowan et al. (2018) found that HRRR precipitation forecasts near mountain ranges in Utah and Lake Tahoe tended to be biased low on the windward slopes and biased high on the leeward slopes, consistent with Colle et al. (2005) simulations over the Wasatch Range using MM5. They point to potential problems with the observations, the model microphysics, or the accurate representation of orographic precipitation processes. Lin et al. (2013) state that “it is a challenge to identify and quantify the sources of precipitation bias.” Their work indicates there are problems with the bulk microphysical parameterizations, as well as issues regarding the freezing level relative to the terrain, with precipitation forecast differences for cold versus warm storms. This is a topic that could be evaluated in the future using PSD freezing-level radar measurements.

Our evaluation metrics are based on point-to-point comparisons, the limitations of which may result in the penalization of the precipitation forecasts if the placement of the precipitation was off by a small distance but the amounts were correct, if the timing of the precipitation was off, or because of the horizontal variability of precipitation in the complex terrain of the western United States. Mass et al. (2002) also point to data representativeness and “...deficiencies in the physical parameterizations of the planetary boundary layer...” as issues that make it difficult for high-resolution models to verify well using traditional point-to-point comparisons, especially in complex terrain. Bytheway and Kummerow (2015) found that “...the areal extent of region within 1 mm h<sup>-1</sup> isohyet is frequently too small...” in the HRRR precipitation rate forecasts, a possible factor in the lower FBS seen in our analysis of HRRR hourly precipitation amounts.

In light of this evaluation of the NOAA WVFT, we propose that in the future we evaluate ways to improve the precipitation forecast currently depicted in the WVFT, such as determining the proper grid points from each model that could be used to provide a neighborhood forecast, accounting for surrounding gradients in the terrain, and presented as a probability of exceedance or percentiles for given precipitation amounts. We could also investigate using the maximum and minimum values from the neighborhood to provide upper and lower boundaries to the QPF, rather than a forecast

from a single grid point. Also, in the future, HRRR time-lagged ensemble forecasts could possibly be used to provide a range of forecast hourly precipitation amounts for the WVFT sites. Until these time-lagged forecasts become available, averaging forecasts from different lead times could be experimented with. In turn, precipitation forecasts from the current and future versions of the RAP and HRRR will have to be evaluated to determine if the improved microphysical packages and data assimilation methods recently implemented have significantly improved the precipitation forecasts, particularly at greater precipitation amounts.

*Acknowledgments.* None of this analysis would have been made possible without the hard work of the NOAA/ESRL/PSD engineering staff who install and maintain the ARO observatories. The authors wish to thank Robert Banta (NOAA/ESRL/CSD), and Mimi Hughes and Francesca Viterbo of NOAA/ESRL/PSD for providing valuable discussions on the analysis methods and editorial comments that much improved the manuscript. We thank Tom Hamill for discussions about the bootstrapping method used for the precipitation ETS and FBS analysis. The authors also wish to thank three anonymous reviewers who provided helpful comments that improved the paper. This research was supported by the Physical Sciences Division of the NOAA/Earth System Research Laboratory.

## APPENDIX

### Precipitation Frequency Bias Scores and Equitable Threat Scores with 95% Confidence Intervals

Figures A1 and A2 show the frequency bias scores (FBS) and equitable threat scores (ETS) for the RAP and HRRR precipitation forecasts, with 95% confidence intervals, to better assess the differences between the forecasts.

## REFERENCES

- Benjamin, S. G., J. M. Brown, and T. G. Smirnova, 2016a: Explicit precipitation-type diagnosis from a model using a mixed-phase bulk cloud-precipitation microphysics parameterization. *Wea. Forecasting*, **31**, 609–619, <https://doi.org/10.1175/WAF-D-15-0136.1>.
- , and Coauthors, 2016b: 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>.
- Bevis, M., S. Businger, T. A. Herring, C. Rocken, R. A. Anthes, and R. H. Ware, 1992: GPS meteorology: Remote sensing of the atmospheric water vapor using the global positioning system. *J. Geophys. Res.*, **97**, 15 787–15 801, <https://doi.org/10.1029/92JD01517>.

- 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>.
- , M. Hughes, K. Mahoney, and R. Cifelli, 2019: A multiscale evaluation of multisensor quantitative precipitation estimates in the Russian River Basin. *J. Hydrometeorol.*, **20**, 447–466, <https://doi.org/10.1175/JHM-D-18-0142.1>.
- Cai, H., and R. E. Dumas 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>.
- Colle, B. A., K. J. Westrick, and C. F. Mass, 1999: Evaluation of MMS and Eta-10 Precipitation Forecasts over the Pacific Northwest during the cool season. *Wea. Forecasting*, **14**, 137–154, [https://doi.org/10.1175/1520-0434\(1999\)014<0137:EOMAEP>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0137:EOMAEP>2.0.CO;2).
- , C. F. Mass, and K. J. Westrick, 2000: MMS precipitation verification over the Pacific Northwest during the 1997–99 cool seasons. *Wea. Forecasting*, **15**, 730–744, [https://doi.org/10.1175/1520-0434\(2000\)015<0730:MPVOTP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0730:MPVOTP>2.0.CO;2).
- , J. B. Wolfe, W. J. Steenburgh, D. E. Kingsmill, J. A. W. Cox, and J. C. Shafer, 2005: High-resolution simulations and microphysical validation of an orographic precipitation event over the Wasatch Mountains during IPEX IOP3. *Mon. Wea. Rev.*, **133**, 2947–2971, <https://doi.org/10.1175/MWR3017.1>.
- Duan, J., and Coauthors, 1996: GPS meteorology: Direct estimation of the absolute value of precipitable water. *J. Appl. Meteorol.*, **35**, 830–838, [https://doi.org/10.1175/1520-0450\(1996\)035<0830:GMDEOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1996)035<0830:GMDEOT>2.0.CO;2).
- 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>.
- Elmore, K. L., H. M. Grams, D. Apps, and H. D. Reeves, 2015: Verifying forecast precipitation type with mPING. *Wea. Forecasting*, **30**, 656–667, <https://doi.org/10.1175/WAF-D-14-00068.1>.
- Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, <https://doi.org/10.1175/WAF-D-17-0144.1>.
- Grams, J. S., W. A. Gallus Jr., S. E. Koch, L. S. Wharton, A. Loughie, and E. E. Ebert, 2006: The use of a modified Ebert–McBride technique to evaluate mesoscale model QPF as a function of convective system morphology during IHOP 2002. *Wea. Forecasting*, **21**, 288–306, <https://doi.org/10.1175/WAF918.1>.
- Gutman, S. I., S. R. Sahm, S. G. Benjamin, B. E. Schwartz, K. L. Holub, J. Q. Stewart, and T. L. Smith, 2004: Rapid retrieval and assimilation of ground based GPS precipitable water observations at the NOAA Forecast Systems Laboratory: Impact on weather forecasts. *J. Meteor. Soc. Japan*, **82**, 351–360, <https://doi.org/10.2151/jmsj.2004.351>.
- 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](https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2).
- Han, F., and I. Szunyogh, 2016: A morphing-based technique for the verification of precipitation forecasts. *Mon. Wea. Rev.*, **144**, 295–313, <https://doi.org/10.1175/MWR-D-15-0172.1>.
- Hwang, Y., A. J. Clark, V. Lakshmanan, and S. E. Koch, 2015: Improved nowcasts by blending extrapolation and model forecasts. *Wea. Forecasting*, **30**, 1201–1217, <https://doi.org/10.1175/WAF-D-15-0057.1>.
- Ikeda, K., M. Steiner, J. Pinto, and C. Alexander, 2013: Evaluation of cold-season precipitation forecasts generated by the hourly updating high-resolution Rapid Refresh model. *Wea. Forecasting*, **28**, 921–939, <https://doi.org/10.1175/WAF-D-12-00085.1>.
- Keyser, D., and R. A. Anthes, 1977: The applicability of a mixed-layer model of the planetary boundary layer to real-data forecasting. *Mon. Wea. Rev.*, **105**, 1351–1371, [https://doi.org/10.1175/1520-0493\(1977\)105<1351:TAOAMM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1977)105<1351:TAOAMM>2.0.CO;2).
- Lin, Y., and K. E. 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](https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm).
- , and B. A. Colle, 2009: The 4–5 December 2001 IMPROVE-2 event: Observed microphysics and comparisons with the Weather Research and Forecasting Model. *Mon. Wea. Rev.*, **137**, 1372–1392, <https://doi.org/10.1175/2008MWR2653.1>.
- , —, and S. E. Yuter, 2013: Impact of moisture flux and freezing level on simulated orographic precipitation errors over the Pacific Northwest. *J. Hydrometeorol.*, **14**, 140–152, <https://doi.org/10.1175/JHM-D-12-019.1>.
- Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, [https://doi.org/10.1175/1520-0477\(2002\)083<0407:DIHRPM>2.3.CO;2](https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2).
- Matrosov, S. Y., F. M. Ralph, P. J. Neiman, and A. B. White, 2014: Quantitative assessment of operational weather radar rainfall estimates over California's Northern Sonoma County using HMT-West data. *J. Hydrometeorol.*, **15**, 393–410, <https://doi.org/10.1175/JHM-D-13-045.1>.
- Neiman, P. J., F. M. Ralph, A. B. White, D. E. Kingsmill, and P. O. G. Persson, 2002: The statistical relationship between upslope flow and rainfall in California's coastal mountains: Observations during CALJET. *Mon. Wea. Rev.*, **130**, 1468–1492, [https://doi.org/10.1175/1520-0493\(2002\)130<1468:TSRBUF>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<1468:TSRBUF>2.0.CO;2).
- , G. A. Wick, F. M. Ralph, B. E. Martner, A. B. White, and D. E. Kingsmill, 2005: Wintertime nonbrightband rain in California and Oregon during CALJET and PACJET: Geographic, interannual, and synoptic variability. *Mon. Wea. Rev.*, **133**, 1199–1223, <https://doi.org/10.1175/MWR2919.1>.
- , F. M. Ralph, A. B. White, D. D. Parrish, J. S. Holloway, and D. L. Bartels, 2006: A multiwinter analysis of channeled flow through a prominent gap along the Northern California Coast during CALJET and PACJET. *Mon. Wea. Rev.*, **134**, 1815–1841, <https://doi.org/10.1175/MWR3148.1>.
- , A. B. White, F. M. Ralph, D. J. Gottas, and S. I. Gutman, 2009: A water vapour flux tool for precipitation forecasting. *Water Manage.*, **162**, 83–94, <https://doi.org/10.1680/wama.2009.162.2.83>.
- , F. M. Ralph, B. J. Moore, and R. J. Zamora, 2014: The regional influence of an intense Sierra barrier jet and landfalling atmospheric river on orographic precipitation in Northern California: A case study. *J. Hydrometeorol.*, **15**, 1419–1439, <https://doi.org/10.1175/JHM-D-13-0183.1>.
- Peckham, S. E., T. G. Smirnova, S. G. Benjamin, J. M. Brown, and J. S. Kenyon, 2016: Implementation of a digital filter initialization in the WRF Model and its application in the Rapid Refresh. *Mon. Wea. Rev.*, **144**, 99–106, <https://doi.org/10.1175/MWR-D-15-0219.1>.
- Pielke, R. A., Sr., 2002: *Mesoscale Meteorological Modeling*. International Geophysics Series, Vol. 78, Academic Press, 676 pp.
- Pinto, J. O., J. A. Grim, and M. Steiner, 2015: Assessment of the High-Resolution Rapid Refresh model's ability to predict mesoscale convective systems using object-based evaluation. *Wea. Forecasting*, **30**, 892–913, <https://doi.org/10.1175/WAF-D-14-00118.1>.

- Ralph, F. M., and Coauthors, 1999: The California Land-falling Jets Experiment (CALJET): Objectives and design of a coastal atmosphere–ocean observing system deployed during a strong El Niño. Preprints, *Third Symp. on Integrated Observing Systems*, Dallas, TX, Amer. Meteor. Soc., 78–81.
- , P. J. Neiman, D. E. Kingsmill, P. O. Persson, A. B. White, E. T. Strem, E. D. Andrews, and R. C. Antweiler, 2003: The impact of a prominent rain shadow on flooding in California’s Santa Cruz Mountains: A CALJET case study and sensitivity to the ENSO cycle. *J. Hydrometeorol.*, **4**, 1243–1264, [https://doi.org/10.1175/1525-7541\(2003\)004<1243:TIOAPR>2.0.CO;2](https://doi.org/10.1175/1525-7541(2003)004<1243:TIOAPR>2.0.CO;2).
- , —, 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](https://doi.org/10.1175/1520-0493(2004)132<1721:SACAOO>2.0.CO;2).
- , —, and R. Rotunno, 2005: Dropsonde observations in low-level jets over the northeastern Pacific Ocean from CALJET-1998 and PACJET-2001: Mean vertical-profile and atmospheric-river characteristics. *Mon. Wea. Rev.*, **133**, 889–910, <https://doi.org/10.1175/MWR2896.1>.
- , E. Sukovich, D. Reynolds, M. Dettlinger, 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. Hydrometeorol.*, **11**, 1286–1304, <https://doi.org/10.1175/2010JHM1232.1>.
- , and Coauthors, 2013: The emergence of weather-related test beds linking research and forecasting operations. *Bull. Amer. Meteor. Soc.*, **94**, 1187–1211, <https://doi.org/10.1175/BAMS-D-12-00080.1>.
- , and Coauthors, 2014: A vision for future observations for western U.S. extreme precipitation and flooding. *J. Contemp. Water Res. Educ.*, **153**, 16–32, <https://doi.org/10.1111/j.1936-704X.2014.03176.x>.
- Waliser, D., and B. Guan, 2017: Extreme winds and precipitation during landfall of atmospheric rivers. *Nat. Geosci.*, **10**, 179–184, <https://doi.org/10.1038/ngeo2894>.
- Weber, B. L., D. B. Wuertz, D. C. Welsh, and R. McPeck, 1993: Quality controls for profiler measurements of winds and RASS temperatures. *J. Atmos. Oceanic Technol.*, **10**, 452–464, [https://doi.org/10.1175/1520-0426\(1993\)010<0452:QCFPMO>2.0.CO;2](https://doi.org/10.1175/1520-0426(1993)010<0452:QCFPMO>2.0.CO;2).
- White, A. B., D. J. Gottas, A. R. Henkel, P. J. Neiman, F. M. Ralph, and S. I. Gutman, 2010: Developing a performance measure for snow-level forecasts. *J. Hydrometeorol.*, **11**, 739–753, <https://doi.org/10.1175/2009JHM1181.1>.
- , and Coauthors, 2012: NOAA’s rapid response to the Howard A. Hanson Dam flood risk management crisis. *Bull. Amer. Meteor. Soc.*, **93**, 189–207, <https://doi.org/10.1175/BAMS-D-11-00103.1>.
- , and Coauthors, 2013: A twenty-first-century California observing network for monitoring extreme weather events. *J. Atmos. Oceanic Technol.*, **30**, 1585–1603, <https://doi.org/10.1175/JTECH-D-12-00217.1>.
- , P. J. Neiman, J. M. Creamean, T. Coleman, F. M. Ralph, and K. A. Prather, 2015: The impacts of California’s San Francisco Bay Area gap on precipitation observed in the Sierra Nevada during HMT and CalWater. *J. Hydrometeorol.*, **16**, 1048–1069, <https://doi.org/10.1175/JHM-D-14-0160.1>.
- Wick, G. A., P. J. Neiman, F. M. Ralph, and T. M. Hamill, 2013: Evaluation of forecasts of the water vapor signature of atmospheric rivers in operational numerical weather prediction models. *Wea. Forecasting*, **28**, 1337–1352, <https://doi.org/10.1175/WAF-D-13-00025.1>.
- Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.
- Willmott, C. J., 1982: Some comments on the evaluation of model performance. *Bull. Amer. Meteor. Soc.*, **63**, 1309–1313, [https://doi.org/10.1175/1520-0477\(1982\)063<1309:SCOTEO>2.0.CO;2](https://doi.org/10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2).
- Zhu, K., Y. Pan, M. Xue, X. Wang, J. S. Whitaker, S. G. Benjamin, and S. S. Weygandt, 2013: A regional GSI-based ensemble Kalman filter data assimilation system for the Rapid Refresh configuration: Testing at reduced resolution. *Mon. Wea. Rev.*, **141**, 4118–4139, <https://doi.org/10.1175/MWR-D-13-00039.1>.