English, J. M., D. D. Turner, D. Dowell, T. I. Alcott, R. Cifelli, and J. L. Bytheway, 2024: Probabilistic Forecasts of Atmospheric River events using the HRRR Ensemble. *J. Operational Meteor.*, **12** (4), 40-53, doi: https://doi.org/10.15191/nwajom.2024.1204.

Journal of Operational Meteorology
Article

# Probabilistic Forecasts of Atmospheric River events using the HRRR Ensemble

*JASON M. ENGLISH*
*NOAA Global Systems Laboratory,*
*Cooperative Institute for Research in Environmental Sciences,*
*University of Colorado, Boulder, CO*

*DAVID D. TURNER, DAVID C. DOWELL, AND TREVOR I. ALCOTT*
*NOAA Global Systems Laboratory, Boulder, CO*

*ROBERT CIFELLI AND JANICE L. BYTHEWAY*
*NOAA Physical Sciences Laboratory, Boulder, CO*

(Manuscript received 9 February 2023; review completed 19 July 2023)

ABSTRACT

**The nine-member High-Resolution Rapid Refresh Ensemble (HRRRE) is evaluated for its ability to forecast five Atmospheric River (AR) events that impacted California in February–March 2019. Two sets of retrospective HRRRE simulations are conducted, a control with the standard set of perturbations (initial and boundary conditions, stochastic parameters, and physics tendency), and an experiment with initial and boundary perturbations only. Reliability plots suggest the HRRRE control represents the observed Stage IV precipitation frequency well at 6-h to 24-h lead times, and rank histograms suggest the ensemble is slightly underdispersive. The HRRRE overpredicts precipitation frequency at the higher (25 mm) threshold. These results suggest the HRRRE is a useful tool to quantify probabilistic forecasts of AR events in this region. Removing stochastic physics perturbations did not substantially impact probabilistic forecasts, suggesting most of the ensemble spread is from initial and boundary condition perturbations. Spatially, ensemble precipitation coefficient of variance is lower (less forecast uncertainty) over the Sierra Nevada range than other regions, suggesting that these ensemble perturbations have a smaller impact on precipitation processes occurring over the Sierra Nevada range. More work should be conducted to understand the impacts of other model perturbations, such as microphysics, on ensemble performance, and to improve Stage IV accuracy with frozen precipitation in mountainous regions.**

## 1. Introduction

Accurate forecasts of the timing, intensity, and location of precipitation and winds can improve preparation for and reduce the negative impacts from Atmospheric River (AR) events, such as flooding, mudslides, and strong winds (Corringham et al. 2019). ARs, long (2000 km or greater) and narrow (300–500 km wide) plumes of enhanced water vapor transport, are a significant source of precipitation for the west coast of North America (Ralph et al. 2004; Dettinger 2013; Lavers et al. 2016), providing up to half of the annual precipitation in California during a typical cold season (Dettinger et al. 2011; Gershunov et al. 2017), recharging local reservoirs/water supply, and reducing drought and wildfire risks. Model forecasts of AR events are often hindered by a complex evolution of synoptic and mesoscale meteorological features (Kingsmill et al. 2006; Ralph et al. 2010; Cannon et al. 2017, 2020) and inadequate observations for model

*Corresponding author address:* Jason M. English, NOAA Global Systems Laboratory, 325 Broadway, Boulder, CO 80305
E-mail: jason.english@noaa.gov

assimilation, particularly over the Pacific Ocean (Stone et al. 2020). Studies demonstrating the value of both deterministic and ensemble modeling systems for AR events have been conducted; however, there have not been any published studies of probabilistic forecasts of AR events over the West CONUS using the High Resolution Rapid Refresh (HRRR) Ensemble (HRRRE; Kalina et al. 2021). As part of the Advanced Quantitative Precipitation Information (AQPI) project (Cifelli et al. 2018, 2022), this work uses the HRRRE to obtain probabilistic information on forecast uncertainty associated with modeling five AR events, continuing the work of English et al. (2021) that used the deterministic HRRR.

The use of high-resolution "convection-permitting" models has been found to improve forecasts of AR events by representing meteorological features that are not resolved by most global models. Martin et al. (2018) compared global and high-resolution models and found high-resolution models to have smaller precipitation errors from ARs due to improved water vapor representation. Huang et al. (2020) studied the impacts of grid spacing in the Weather Research and Forecasting (WRF; Skamarock et al. 2019) model on accuracy of AR-related precipitation extremes and found improvements as large as 40–60% in the fine scale (3-km) version relative to coarse-scale (27-km) simulations. Gowan et al. (2018) found several high-resolution models—the High-Resolution Rapid Refresh (HRRR; James et al. 2022; Dowell et al. 2022), the North American Model (NAM; Mathiesen and Kleissl 2011) 3-km nest, and the National Center for Atmospheric Research (NCAR) Ensemble—to be more accurate than coarser operational models when comparing Quantitative Precipitation Forecasts (QPF) to Quantitative Precipitation Estimates (QPE) over the western contiguous United States (CONUS) during the cool season. Two intercomparison studies have found the HRRR to perform the best for AR events among high-resolution deterministic models (Gowan et al., 2018; Dougherty et al. 2021). Additionally, English et al. (2021) found HRRRv4 to demonstrate improved forecasts of AR events relative to HRRRv3, but both model versions exhibit QPF dry biases in the San Francisco Bay Area and along the Pacific Coast and QPF wet biases in the Sierra Nevada range.

In recent years, model ensembles have become a useful tool to help quantify atmospheric forecast uncertainty inherent with numerical weather prediction (NWP) (Zhang and Pu 2010). Instead of making a single (deterministic) forecast of the most likely future state of the atmosphere, a set (or ensemble) of forecasts is produced, to give an indication of the range of possible future states of the atmosphere. Often, specific parameter(s) are perturbed, such as initial/boundary conditions or physics, to produce the output of the individual ensemble members. Many studies have found model ensembles to be more accurate than their deterministic counterparts in general (Atger 2001; Grimit and Mass, 2002; Rodwell 2006; Vokoun and Hanel 2018; Zhao et al. 2020), and several studies of model ensemble forecasts of ARs impinging on the West CONUS have been conducted. Yuan et al. (2005) evaluated probabilistic forecasts from the National Centers for Environmental Prediction (NCEP) Regional Spectral Model (RSM; Juang and Kanamitsu 1994; Juang et al. 1997) over the Southwest CONUS, and found a general wet bias in the ensembles, and concluded that the value of probabilistic forecasts depends strongly on geography, threshold, and reference dataset. Yuan et al. (2008) compared QPF from several different time-lagged ensemble models to Stage IV QPE in the Northern California cool season and found model choice, model physics, and initial and boundary conditions to all impact forecast uncertainty. Peel and Wilson (2008) found the Canadian Ensemble Forecast System (CEFS) to perform better in the cool season than the warm season at higher thresholds, but that uncertainty in QPE accuracy is much higher in the cool season due to the impact of snow events. McColor and Stull (2008) found that QPF accuracy during the British Columbia cool season was improved in the Geophysical Disaster Computational Fluid Dynamics Centre (GDCFDC) real-time suite when including lower-resolution models in the ensemble, compared to including only high-resolution models. Brown et al. (2012) evaluated QPF from NCEP's Short-Range Ensemble Forecast (SREF; Du et al. 2009) system over the Sierra Nevada region and found its ensembles to overestimate light precipitation and underestimate heavy precipitation versus Stage IV, and to be overconfident (underdispersive). Wick et al. (2013) evaluated ARs impacting the West CONUS with five operational ensemble forecast systems. Although AR occurrences were forecasted with a 10-day lead time, skill degraded with increasing lead time, with an average error of more than 800 km at a 10-day lead time, and a 1–2 deg southward position bias at a 7-day lead time. Although the ensemble was able to forecast timing, the location was more difficult for the model to capture. Brown et al. (2014) evaluated precipitation,

temperature, and streamflow from the Hydrologic Ensemble Forecast Service (HEFS; Seo et al. 2010; Demargne et al. 2010, 2014).) across a 20-year period over the California-Nevada River Forecast Center (CNRFC) and found reasonably accurate forecasts during the cold season, but an underestimation during the highest precipitation events. Lewis et al. (2017) compared Global Ensemble Forecast System (GEFS; Hamill et al. 2013) QPF to Snow Telemetry (SNOTEL) stations over the West CONUS during the 2013–2015 cool seasons and found widespread dry biases, with low ensemble reliability and insufficient spread.

Because the deterministic HRRR has demonstrated accurate QPF relative to other high-resolution models, and ensemble models are often more accurate than their deterministic counterparts, our hypothesis is that HRRRE probabilistic forecasts of AR events should produce reliable QPF probability curves relative to other ensemble models. Regardless, the contributions of various types of perturbations on the mean and spread of ensemble forecasts of AR events impacting the West CONUS have not been systematically evaluated. The HRRRE includes perturbations to initial and boundary conditions (the most common types of perturbations), and stochastic physics perturbations. Our hypothesis is that initial and boundary conditions may be the primary source of ensemble spread/forecast uncertainty, due in part to fewer observations over the Pacific Ocean. Conversely, stochastic physics perturbations, which were identical to those in previous HRRRE experiments (Kalina et al. 2021), may produce a relatively smaller contribution to ensemble spread, due to the presence of strong synoptic and orographic forcing for ascent and the largely non-convective nature of AR events. And finally, it is unclear how much ensemble spread varies spatially over California during AR events. Our hypothesis is that ensemble spread may be larger along the Pacific Coast, where numerous complex precipitation processes occur, and smaller over the Sierra Nevada range, where much of the precipitation is driven by large-scale orography.

In this paper we explore three questions: 1) How accurate are HRRRE probabilistic forecasts of AR events over the West CONUS? 2) What are the impacts of stochastic physics perturbations on ensemble accuracy and spread? 3) How does precipitation bias and ensemble spread vary spatially? We will answer these questions via the following analyses, respectively: 1) Evaluate HRRRE Reliability plots and Rank histograms, and compare ensemble mean spatial maps to the deterministic HRRR; 2) compare a set of ensemble runs with and without stochastic physics and quantify QPF differences between them; and 3) evaluate spatial maps of ensemble spread.

## 2. Data and methods

### *a. The HRRR Ensemble*

The HRRRE is a prototype ensemble modeling system that uses a single physics package/single dynamic core with stochastic perturbations to the physical parameterizations and initial and boundary condition perturbations to produce an ensemble forecast (Kalina et al. 2021). The HRRRE contains a 36-member ensemble analysis system and a 9-member forecast system producing 36-h forecasts, although we produce 24-h forecasts in this study because we focus our investigation on lead times up to 24-h. The HRRRE encompasses an area slightly larger than the CONUS with a convection-allowing 3-km horizontal grid spacing. The HRRRE uses the Advanced Research version of the Weather Research and Forecasting (WRF-ARW) dynamic core and the Rapid Refresh/ High Resolution Rapid Refresh (RAP/ HRRR) physics suite (Benjamin et al. 2016; Olson et al. 2019; James et al. 2022; Dowell et al. 2022). The HRRRE includes four types of perturbations: initial condition (IC) perturbations, boundary condition (BC) perturbations, stochastic parameter perturbations (SPP) (Palmer 2001), and stochastic perturbations of physics tendencies (SPPT) (Buizza et al. 1999). The SPP scheme in HRRRE consists of a random pattern generator that creates a vertically uniform perturbation field with prescribed spatiotemporal correlations, and is applied to numerous parameters in numerous physics schemes (Kalina et al. 2021). The SPPT scheme perturbs the total physics tendencies for temperature, humidity, and wind. Because the HRRRE uses a single dynamic core and a single physics suite, ensemble spread may be more limited than multi-dynamic or multi-physics ensemble systems, but this HRRRE design enables statistically consistent ensemble distribution (Bowler et al. 2009; Berner et al. 2009; Sanchez et al. 2015). More details on the HRRRE configuration, including the complete list of SPP fields and the magnitudes of their perturbations, are provided in Kalina et al. 2021. In this study, we use a temporal scale of 72 h instead of 6 h for SPP and SPPT perturbations, based on more recent studies at the National Oceanic and Atmospheric Administration

**Table 1**. Time Periods of AR events studied. Times listed below are valid times; hence model initialization times will differ by each lead time. For all lead time evaluations, model cycles are initialized at 00 and 12 UTC each day; hence, five to six initialization times are utilized for each AR event, depending on lead time.

| AR Event (2019) | Valid Time Period Averaged |
|---|---|
| 2-4 Feb | 12 UTC 02-Feb to 12 UTC 04-Feb |
| 13-15 Feb | 06 UTC 13-Feb to 06 UTC 15-Feb |
| 25-27 Feb | 06 UTC 25-Feb to 06 UTC 27-Feb |
| 2-4 Mar | 06 UTC 02-Mar to 06 UTC 04-Mar |
| 5-7 Mar | 06 UTC 05-Mar to 00 UTC 07-Mar |

(NOAA) Global Systems Lab (GSL) that resulted in improved ensemble spread and reduced error of near-surface meteorological fields (Isidora Jankov, personal communication). The HRRRE was run in "real-time" at NOAA GSL from 2017 through 2021, and evaluations of its configuration and verification are informing the next-generation ensemble model development, which will be based on the new Rapid Refresh Forecast System (RRFS) instead of the HRRR.

*b. Experimental design*

We set up and run two continuous HRRRE retrospective simulations from 1 February 2019 through 10 March 2019: a HRRRE control with IC, BC, SPP, and SPPT ensemble perturbations (using the same configuration as the "real-time" HRRRE run at NOAA GSL), and a HRRRE experiment with IC and BC perturbations only. This time period encompasses five AR events that occurred in California (Table 1). Comparing the two simulations can provide insight regarding how much SPP and SPPT perturbations impact forecasts of AR events over the West CONUS. The HRRRE is initialized twice a day at 00 and 12 UTC using soil state from a companion deterministic HRRRv4 retrospective simulation, and 24-h forecasts are produced. More details on the deterministic

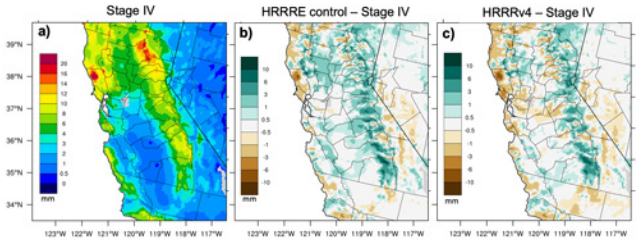


**Figure 1**. Spatial maps of average 6-h accumulation (mm), averaged across the 36-h or 48-h peak precipitation time periods for all five AR events (Table 1). a) Stage IV. b) HRRRE control (IC/BC/SPP/SPPT) bias (6-h lead time). c) Deterministic HRRRv4 (6-h lead time) bias. HRRRE bias is calculated as HRRRE ensemble average minus Stage IV; HRRRv4 bias is calculated as HRRRv4 minus - Stage IV; blue-green colors are a model wet bias and brown colors are a dry bias. Comparisons are "matching"; hence, the three datasets include initialization times 00 and 12 UTC (valid times 06 and 18 UTC) each day. *Click image for an external version; this applies to all figures hereafter.*

HRRRv4 retrospective simulation and the five AR events are provided in English et al. (2021).

The choice of evaluation metrics and corresponding spatiotemporal domains to average is complex. Although evaluating over smaller domains or time periods or across individual AR events may facilitate identifying differences between the control and the experiment, differences are often not statistically significant when looking at individual weather events or specific geographic locations. Indeed, in English et al. (2021), we found that although several AR events had some interesting unique characteristics, differences between two model runs were not statistically significant when evaluating any individual event, and we needed to average across all five AR events to observe significant differences between two model runs. Hence, we follow a similar approach for this study: We evaluate QPF across the time period in which the most significant precipitation occurred for each AR event (generally 36-h or 48-h in length per event; Table 1) over our designated AQPI domain (33.38–41.48N, 118.28–123.88W; Fig. 1), and average across all five AR events. We remove a small region near a recently identified faulty rain gauge, as discussed further in the next paragraph. We evaluate ensemble forecast performance via reliability plots (Wilks 1995), rank histograms (Anderson 1996; Hamill and Colucci 1997; Talagrand et al. 1997; Wilks 2019), frequency bias plots, fractions skill score (FSS) (Roberts and Lean 2008), and precipitation bias. Model

QPF is compared to Stage IV QPE. Stage IV is an hourly/6-hourly product produced by the twelve river forecast centers (Lin and Mitchell 2005; Nelson et al. 2016). Over the AQPI evaluation domain, Stage IV is produced by the CNRFC, and is based solely on gauges and climatology (i.e., without using radar-derived precipitation estimates, due to radar gaps/blockages across many parts of California). Though Stage IV has some known weaknesses, often associated with mountainous regions and frozen precipitation (Nelson et al. 2016; Herman and Schumacher 2018), it is generally regarded as a high-quality, gridded multisensory QPE product, and is often chosen as the reference dataset when evaluating other QPE products (Wu et al. 2012; Gourley et al. 2010; Lin and Hou 2012). However, all QPE products including Stage IV are known to have uncertainties in parts of the CNRFC region, particularly in mountainous terrain and at temperatures below freezing (Peel and Wilson 2008; Smalley et al. 2014; Lundquist et al. 2019; Bytheway et al. 2020; English et al. 2021). To address these uncertainties, we also average ensemble performance over two different altitude domains: the AQPI domain at elevations below 1500 m (eliminating most of the domain that involves mountainous terrain and temperatures below freezing, where Stage IV is least reliable), and the AQPI domain at elevations above 1500 m. Additionally, we remove a small region (37.9–38.1 N and 121.4–121.6 W) from our computations for Reliability plots and Rank histograms to account for a faulty gauge recently identified in the Russian River basin. This gauge (Venado) was improperly sited and producing accumulations that were much too high compared to a co-located weighing type gauge (Andrew Martin, personal communication).

## 3. Analysis and discussion

### *a. Ensemble mean spatial precipitation*

While the spatial distribution of precipitation varies considerably across the five AR events (see English et al. 2021), mean 6-h precipitation is generally highest along the Pacific Coast and the Sierra Nevada range, and lowest in the Central Valley, (Fig. 1a). Because of orographic enhancement, the heaviest precipitation occurs in the mountainous regions, particularly the coastal mountains north of the San Francisco Bay Area and the Sierra Nevada range north of Lake Tahoe, where mean precipitation rates reach roughly 20 mm (6 h)$^{-1}$ (160 mm per 48-h AR event). HRRRE control mean precipitation compares reasonably well to Stage IV (generally within 3 mm (6 h) $^{-1}$), but is drier along much of the Pacific Coast, especially in the mountains north of the San Francisco Bay Area (however, the recently identified faulty Venado gauge is in this region and likely to have been included in the production of the Stage IV QPE), and wetter than Stage IV to the east, particularly in the Sierra Nevada range (Fig. 1b).

The HRRRE control mean has a somewhat similar bias map to the deterministic HRRRv4 (Fig. 1c), which shares the same physics as the HRRRE configuration studied here, although the HRRRE control dry bias is reduced, particularly in the San Francisco Bay Area and along the Pacific Coast. The deterministic HRRRv4 dry biases versus Stage IV in the San Francisco Bay Area

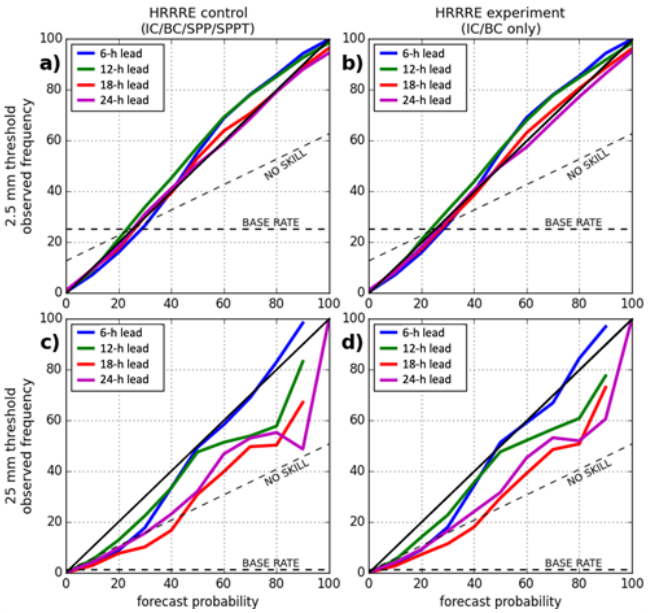


**Figure 2**. Reliability plots of 6-h accumulated precipitation for the HRRRE control (IC/BC/SPP/SPPT) and HRRRE experiment (IC/BC only) at two thresholds (2.5 mm and 25 mm) and four lead times (6-h, 12-h, 18-h, 24-h) compared to Stage IV. Data include the 36-h to 48-h time period in which the most significant precipitation for each AR event (Table 1) across the designated AQPI domain (33.38–41.48N, 118.28–123.88W), with a small domain surrounding the faulty Venado gauge removed (37.9–38.1N, 121.4–121.6W). The "no skill" lines indicate where reliability and resolution are equal and therefore the forecast has zero skill relative to a reference forecast (i.e., random chance). Comparisons are "matching"; hence, the datasets include any valid times corresponding to the respective lead times from the 00 and 12 UTC HRRRE initialization times each day.

and along the Pacific Coast are possibly related to errors in the modeled temperature profile and/or the vertical distribution of water vapor, while integrated water vapor (IWV), wind speed, and wind direction all compared well to available observations at nearby Atmospheric River Observatories (English et al. 2021), and were found in studies with the deterministic HRRRv3 as well (Darby et al. 2019; Dougherty et al. 2021). Dougherty et al. (2021) suggested that HRRR wind direction may be responsible for the HRRR QPF dry bias along the Pacific Coast, as 12-h forecasts from the HRRRv3 had some biases in wind direction compared to nearby ARO stations during the 2018/2019 cool season. The HRRRv4 wet biases versus Stage IV in the Sierra Nevada range are at least partly attributed to errors and uncertainties with detecting frozen precipitation in QPE products such as Stage IV; when constrained to liquid precipitation only, HRRRv4 and Stage IV have excellent agreement (English et. al 2021). The HRRRE experiment mean bias map looks visually similar to the HRRRE control (not shown), and differences between the control and experiment are discussed in more depth in the following sections.

*b. Reliability plots*

The HRRRE is compared to Stage IV QPE via reliability plots of 6-h accumulation at two precipitation thresholds (Fig. 2). A 25-km Gaussian smoother is applied to the probability of exceedance forecasts to reduce the impact of small location errors. We chose 25-km to be consistent with the operational High Resolution Ensemble Forecast system version 2 (HREFv2). A perfect ensemble would match the observed frequency at each probability, following the black diagonal line. At a 2.5-mm threshold (Fig. 2a), HRRRE control forecast probabilities generally agree well with Stage IV frequency, particularly at 18-h and 24-h lead times, where they differ by <5%. At shorter lead times, HRRRE control forecast probabilities are slightly lower than Stage IV frequencies, particularly at probabilities >50%, suggesting the model slightly underpredicts the probability of precipitation exceeding the given thresholds. At a 25-mm threshold, HRRRE control 6-h forecasts agree well with Stage IV frequency at probabilities above 50%, but overpredict precipitation frequency at lower probabilities. At lead times >6 h, the HRRRE control more consistently overpredicts the probability of precipitation, and this overprediction generally increases with longer lead times, except for

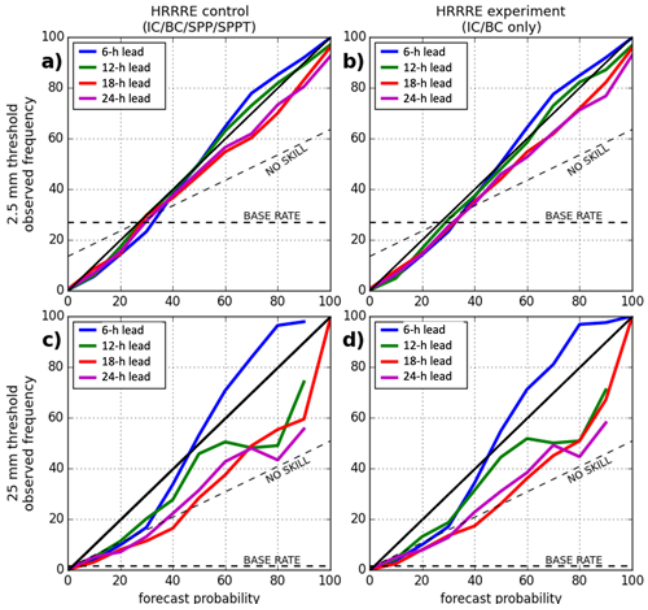


**Figure 3.** Same as Figure 2, but including only grid boxes at <1500 m elevation in the AQPI domain.

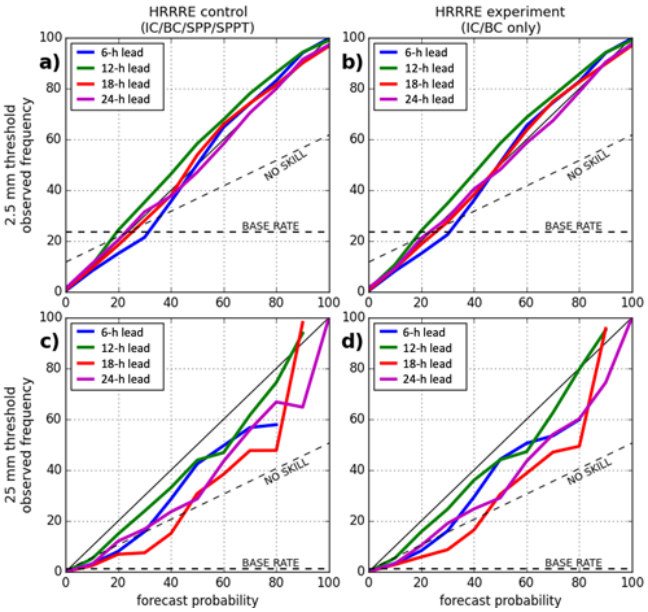


**Figure 4.** Same as Figure 2, but including only grid boxes at >1500 m elevation in the AQPI domain.

probabilities from about 55 to 80% where 18-h lead times overpredict more than 24-h lead times.

Comparing the HRRRE control to the HRRRE experiment can help answer the question: What are the impacts of SPP and SPPT perturbations on ensemble accuracy and spread? The difference in performance metrics between the HRRRE control and experiment provide an understanding of the contributions of SPP

and SPPT to ensemble QPF mean and/or spread, and therefore the importance of these perturbations on model uncertainty and forecast probabilities for AR events impacting the West CONUS. Reliability plots for the HRRRE experiment (Figs. 2b, 2d) generally differ from the HRRRE control by <5% at a given observed frequency (Figs 2a, 2c), suggesting that SPP and SPPT perturbations have small contributions to the probability distributions.

Next, we produce reliability plots for the low (<1500 m) and high (>1500 m) altitude domains (Figs. 3 and 4, respectively). In the low altitude (<1500 m) domain at a 2.5 mm threshold, HRRRE control forecast probabilities increase relative to Stage IV frequency when including only data below 1500 m than when evaluating over the full AQPI domain (Fig. 3a versus Fig. 2a, respectively). At forecast probabilities >50%, this translates to improved ensemble performance at shorter (6-h and 12-h) lead times, and degraded performance at longer (18-h and 24-h) lead times. At forecast probabilities <50%, this translates to a slight overprediction of precipitation probability relative to Stage IV. In the low altitude (<1500 m) domain at a 25 mm threshold, the HRRRE overforecasts precipitation probability at lead times 12 h and longer, whereas HRRRE underpredicts precipitation probability at shorter (6-h) lead times (Fig. 3c). Forecast probabilities in the high altitude (>1500 m) domain are generally <5% different than probabilities over the full AQPI domain at both 2.5 and 25 mm thresholds, except for shorter (6-h) lead times at a 25 mm threshold, where HRRRE control forecast probabilities increase relative to Stage IV in the high altitude (>1500 m) domain (Fig. 4c versus Fig. 2c, respectively). As with the evaluation over the full AQPI domain, reliability plots for the HRRRE experiment are similar to the HRRRE control (generally differ by <5%) over the low altitude (Fig. 3) and high altitude (Fig. 4) domains as well.

### *c. Rank histograms*

Rank histograms are provided at two lead times (6-h and 24-h) to show the relative ranking of the observed precipitation among the HRRRE members (Fig. 5). A perfect ensemble would have the same relative frequency across all ranks, meaning the observation is indistinguishable among the model ensemble members, suggesting a consistent degree of ensemble dispersion. At 6-h lead times, Stage IV falls outside the minimum and maximum ensemble members

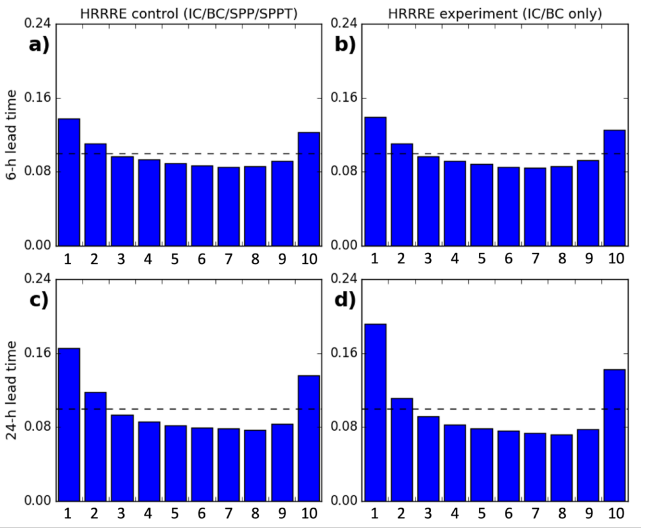


**Figure 5.** Rank Histograms of 6-h accumulated precipitation for the HRRRE control (IC/BC/SPP/SPPT) and HRRRE experiment (IC/BC only) at two lead times (6-h and 24-h). The histograms report the frequency of occurrence of Stage IV relative to the nine HRRRE members. The dashed line represents a flat histogram, where Stage IV has a frequency of occurrence that is statistically indistinguishable among the nine HRRRE members (hence, a total of 10 bins) (Wilks 2019).

slightly more often than expected (about 13–14% rather than 10% occurrence), suggesting HRRRE is slightly underdispersive (Fig. 5a). The underdispersiveness is slightly larger at longer (24-h) lead times (Fig. 5c), suggesting the ensemble spread does not sufficiently capture forecast uncertainty at longer lead times. When evaluating longer (24-h) lead times as a function of elevation, the HRRRE is less underdispersive over the low elevation (<1500 m) domain (Figs. 6a), and more underdispersive over the high elevation (>1500 m) domain (Figs. 6c). However, because of uncertainty with Stage IV at high elevation, it is unclear whether the HRRRE is not accurately capturing the precipitation spread, or if there are errors with Stage IV. Indeed, the small-scale maxima and minima in the HRRRE mean precipitation fields are physically consistent with orographic processes, and may not be adequately included in the Stage IV data. Regardless, the HRRRE was found to be underdispersive in other applications as well, such as warm-season convective events (Grim et al. 2022). This underdispersiveness is a well-known deficiency with model ensembles in general (Berner et al. 2017), and is consistently noted in other ensemble evaluations of AR events impacting the West CONUS

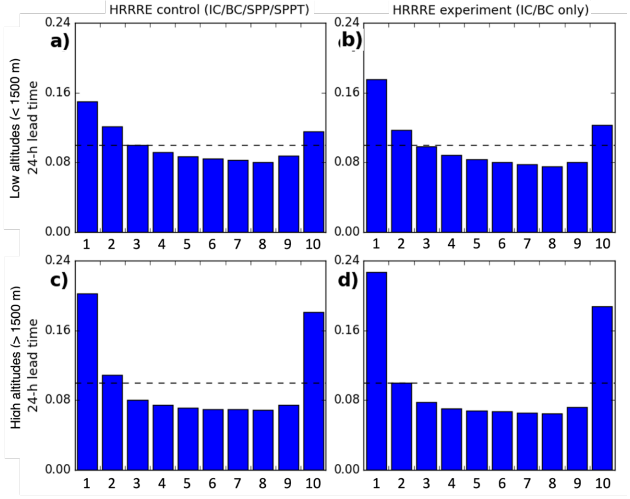


**Figure 6.** Rank Histograms of 6-h accumulated precipitation for the HRRRE control (IC/BC/SPP/SPPT) across two elevation domains at 24-h lead time. The histograms report the frequency of occurrence of Stage IV relative to the nine HRRRE ensemble members. The dashed line represents a flat histogram, where Stage IV has a frequency of occurrence that is statistically indistinguishable among the nine HRRRE members (hence, a total of 10 bins) (Wilks 2019).

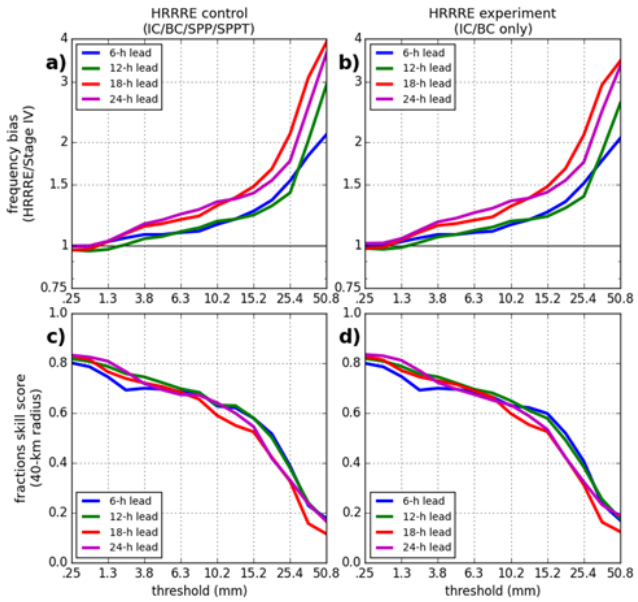


**Figure 7.** Frequency Bias and Fractions Skill Score of 6-h accumulated precipitation for all of the ensemble members in the HRRRE control (IC/BC/SPP/SPPT) and HRRRE experiment (IC/BC only) at a range of thresholds and four lead times (6-h, 12-h, 18-h, 24-h) compared to Stage IV. Data include the 36-h to 48-h time period in which the most significant precipitation for each AR event (Table 1) across the designated AQPI domain (33.38–41.48N, 118.28–123.88W), with a small domain surrounding the faulty Venado gauge removed (37.9–38.1N, 121.4–121.6W). Comparisons are "matching"; hence, the datasets include any valid times corresponding to the respective lead times from the 00 and 12 UTC HRRRE initialization times each day.

(McCollor and Stull 2008; Yuan et al. 2008; Peel and Wilson 2008; Lewis et al. 2017).

At 6-h lead times, rank histograms are similar between the HRRRE control (Fig. 5a) and HRRRE experiment (Fig. 5b); the frequency distributions differ by generally <2% for each ensemble member. At 24-h lead times, the HRRRE experiment (Fig. 5d) is slightly more underdispersive than the HRRRE control (Fig. 5c). This is true both at low altitudes (Fig. 6b versus Fig. 6a, respectively) and high altitudes (Fig. 6d versus Fig. 6c, respectively). This suggests that at longer (24-h) lead times, the inclusion of SPP and SPPT perturbations help increase ensemble spread, and therefore, provide a more accurate representation of forecast uncertainty.

### *d. Frequency bias and fractions skill scores (FSS)*

HRRRE frequency bias plots show the HRRRE control to have an excellent comparison to Stage IV at thresholds less than 1.3 mm at all lead times (frequency bias is between 0.98 and 1.05) (Fig. 7a). At larger thresholds, the HRRRE overforecasts the frequency of precipitation (frequency bias greater than one). Frequency bias increases at longer lead times at the larger thresholds. These results differ from Brown et al. (2012), where they found the SREF to underestimate heavy precipitation. However, there are many differences between the two ensemble systems; the HRRRE leverages an hourly cycled data assimilation ensemble for model spin up, and is run at convection-permitting grid spacing and generates a more complex and amplified 3-D vertical velocity field, particularly when representing orographic forcing and convective updrafts. The FSS is highest (approximately 0.8) at smallest thresholds, and decreases to less than0.2 at the largest thresholds (Fig. 7c). In contrast to frequency bias, FSS does not change much at longer lead times (generally within 10% at a given threshold).

Frequency bias differs by less than 3% between the HRRRE control (Fig. 7a) and HRRRE experiment at thresholds less than about 30 mm (Fig. 7b). At thresholds greater than about 30 mm, the HRRRE experiment has a larger frequency bias relative to the control at lead

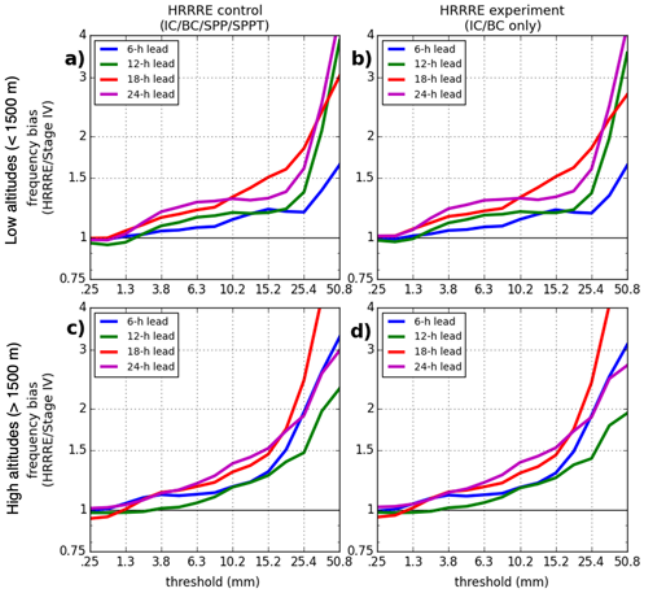


**Figure 8.** Frequency Bias of 6-h accumulated precipitation for all of the ensemble members in the HRRRE control (IC/BC/SPP/SPPT) and HRRRE experiment (IC/BC only) at a range of thresholds and four lead times (6-h, 12-h, 18-h, 24-h) compared to Stage IV. Data include the 36-h to 48-h time period in which the most significant precipitation for each AR event (Table 1) across the two elevation domains within the designated AQPI domain (33.38–41.48N, 118.28–123.88W).

times 12 h and longer. FSS as a function of threshold are also similar between the HRRRE control (Fig. 7c) and the HRRRE experiment (Fig. 7d) at all lead times, generally differing by less than 3%.

Next, we produce frequency bias plots for the low (less than 1500 m) and high (greater than 1500 m) altitude domains (Fig. 8). Frequency bias is larger over the high-altitude domain, particularly at shorter (6-h) lead times at thresholds greater than 15 mm. In particular, for 6-h lead times at a threshold of 25.4 mm, frequency bias is 1.2 over the low altitude domain (Fig. 8a) but is 2.0 over the high-altitude domain (Fig. 8c) (67% higher). As with the evaluation over the full domain, frequency bias differs by less than3% between the HRRRE control and HRRRE experiment at thresholds less than about 30 mm over both the low altitude domain (Fig. 8b versus Fig. 8a, respectively) and the high-altitude domain (Fig. 8d versus Fig. 8c, respectively).

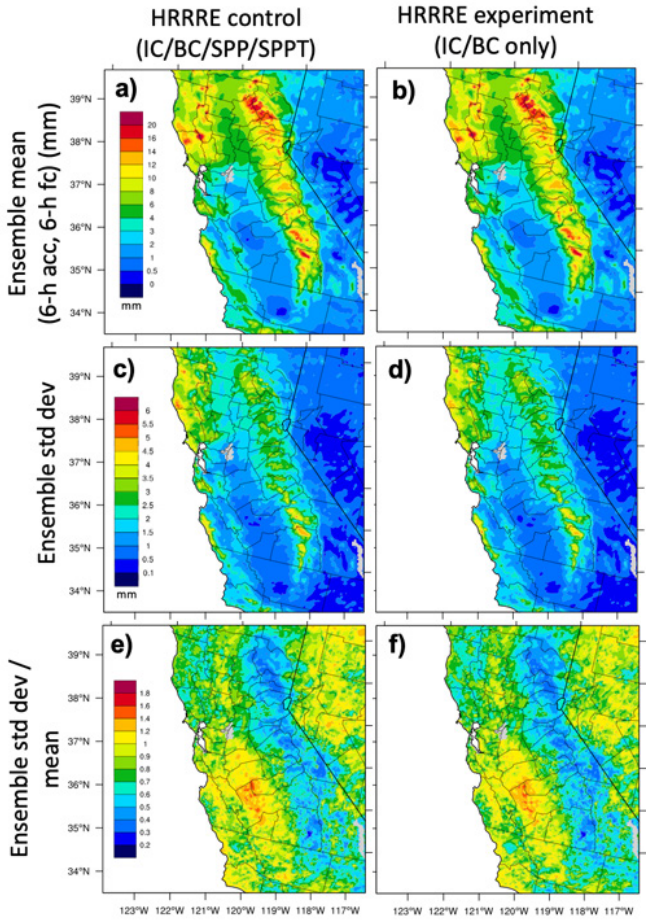


**Figure 9.** Spatial maps of 6-h accumulation (mm) ensemble mean and standard deviation, averaged across the 36-h or 48-h peak precipitation time periods for all five AR events (Table 1) for the HRRRE control (IC/BC/SPP/SPPT) and HRRRE Experiment (IC/BC only) (6-h lead times). (a,b) ensemble mean (mm). (c,d) ensemble standard deviation (mm). (e,f) normalized ensemble standard deviation (standard deviation divided by ensemble mean). Comparisons are "matching"; hence, the datasets include HRRRE initialization times 00 and 12 UTC (Stage IV valid times 06 and 18 UTC) each day.

### *e. Spatial variation of ensemble spread*

Spatial maps of 6-h accumulated ensemble mean and standard deviation for the HRRRE control (6-h lead times) illustrate the spatial distribution of ensemble spread (Fig. 9). Ensemble standard deviation is larger along the Pacific Coast and in the Sierra Nevada range (Fig. 9c). However, the standard deviation is strongly related to the mean value, which varies significantly and also peaks along the Pacific Coast and in the Sierra Nevada range (Fig. 9a). When calculating the

coefficient of variation (dividing standard deviation by the mean precipitation value), ensemble spread per unit accumulated precipitation is largest in the Central Valley, and lowest in the Sierra Nevada range (Fig. 9e). The larger normalized ensemble spread in the Central Valley is related to the small values of accumulated precipitation (generally <1 mm per 6 h) in the denominator. The lower normalized ensemble spread in the Sierra Nevada range suggests that the ensemble perturbations (IC, BC, SPP, and SPPT) occurring in this HRRRE model configuration do not affect precipitation processes in the Sierra Nevada range as much as in other geographic regions. This suggests that the processes responsible for precipitation over the Sierra Nevada range (orographic forcing and microphysical processes) are not being perturbed as much. However, some of the regional differences in coefficient of variation are due to biases in mean precipitation. Over the Sierra Nevada range, the HRRRE has a mean wet bias of roughly 20% (Fig. 1b), which contributes to a 20% reduction in coefficient of variation. Therefore, adjusting to a mean bias of zero would increase the coefficient of variation over the Sierra Nevada range by about 20% (from roughly 0.4 to about 0.6). Likewise, along the Pacific coast, the HRRRE has a mean dry bias of about 10%, which contributes to a 10% increase in coefficient of variation. Adjusting to a mean bias of zero would decrease the coefficient of variation over the Pacific coast by about 10% (from roughly 0.9 to 0.8). This reduces the regional differences, but the ensemble spread over the Sierra Nevada range remains generally less than other geographic regions, with the possible exception of the mountains north of the San Francisco Bay Area (recall the recently identified faulty Venado gauge in this region). There is some uncertainty in this assessment as Stage IV is less reliable over the Sierra Nevada range. Conversely, the relatively higher ensemble spread over the Coastal Range suggests that the ensemble perturbations do have a significant impact on precipitation in this region. Spatial maps of the HRRRE control ensemble coefficient of variation (Fig. 9e) appear visually similar to the HRRRE experiment (Fig. 9f), suggesting that SPP and SPPT perturbations do not meaningfully impact the spatial biases in any particular location.

## 4. Conclusions

The results presented here are a first evaluation of precipitation forecasts of AR events from the HRRRE. Probabilistic forecasts from the HRRRE for five AR events that impacted California in February and March 2019 were investigated to answer three science questions:

1. How accurate are HRRRE probabilistic forecasts of AR events over the West CONUS? The ensemble metrics presented here (reliability plots and rank histograms) suggest that the forecast probability from the HRRRE with its current real-time perturbations (IC/BC/SPP/SPPT) provide a reasonably accurate representation of forecast uncertainty of AR events impacting central California at 6-h to 24-h lead times. Reliability plots (Fig. 2) suggest that the ensemble members of the HRRRE control have a reliable representation of observed frequency across all lead times studied (6-h to 24-h) at the 2.5-mm threshold. The HRRRE overpredicts the frequency of precipitation occurrence at the 25-mm threshold. Rank histograms suggest the HRRRE spread to be slightly underdispersive, which is a common trait of many model ensemble systems. The HRRRE is less underdispersive at elevations below 1500 m, where Stage IV is known to be most reliable, with higher density of precipitation gauges, less climatological adjustment, and less frozen precipitation.

2. What are the impacts of stochastic physics perturbations on ensemble accuracy and spread? The differences between the HRRRE control and HRRRE experiment are small by most metrics investigated here (reliability plots, rank histograms, frequency bias, FSS, and spatial maps of ensemble mean precipitation and standard deviation). This suggests that SPP and SPPT perturbations make relatively small contributions to ensemble accuracy of precipitation forecasts from AR events, which is supported by spatial maps of ensemble mean precipitation (Fig. 9) that show small differences between the HRRRE control and HRRRE experiment. One exception is that rank histograms of the HRRRE experiment are slightly less underdispersive at 24-h lead time, suggesting that SPP and SPPT perturbations contribute meaningfully to an increase of ensemble spread at longer lead times.

Overall, IC/BC perturbations make a much larger contribution to ensemble performance/spread than SPP and SPPT perturbations.

3. How does precipitation bias and ensemble spread vary spatially? Spatial maps of average precipitation (Fig. 1) suggest the HRRRE mean reasonably captures the observed spatial distribution of precipitation, but is wetter than Stage IV over the Sierra Nevada range, and drier than Stage IV along the coastal range north of the San Francisco Bay Area (however, Stage IV may not be reliable in this region due to the recently identified faulty Venado gauge). Spatial maps of ensemble coefficient of variation show that ensemble spread is largest over the Central Valley, and smallest over the Sierra Nevada range, although a HRRRE precipitation wet bias contributes to part of the reduction over the Sierra Nevada range. These differences in coefficient of variation suggest there is greater uncertainty in forecasts over the Central Valley, and less uncertainty over the Sierra Nevada range. This suggests that the HRRRE ensemble perturbations (IC/BC/SPP/SPPT) have a small effect on precipitation forecast uncertainty over the Sierra Nevada range, and that other model parameters/uncertainties are likely at play. Rank histograms show the HRRRE to be more underdispersive above 1500 m elevation, suggesting the ensemble isn't capturing sufficient forecast uncertainty, although Stage IV is less reliable at higher elevations.

These results demonstrate the value and accuracy of probabilistic forecasts of precipitation from the single dynamic core/single physics suite used in HRRRE during AR events impacting the West CONUS, which provides useful guidance towards developing the next-generation RRFS deterministic model and ensembles based on it. Future work exploring perturbations to other parameters that might be relevant to the meteorological processes present during AR events would be useful. For example, Jeworrek et al. (2021) investigated the impacts of changing physics, microphysics, and grid spacing specifications in the WRF model (on which the HRRR is based) for a year of precipitation forecasts in British Columbia and concluded that the choice of cumulus and microphysics parameterizations had the largest impact on precipitation forecasts. Further work to improve the reliability of QPE products at high elevations when frozen precipitation is present, such as a probabilistic QPE product (Bytheway et al. 2022) would be useful to better quantify observed precipitation as well as more confidently evaluate model QPF accuracy.

*Acknowledgments*. Helpful comments from a GSL internal reviewer (Evan Kalina) and two peer reviewers improved this manuscript. This work is funded in part by the AQPI research program via NOAA Award 3RR2NAQ-P02 and by the NOAA Cooperative Agreement with CIRES, NA17OAR4320101.

## REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530, CrossRef.

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes in Geophysics*, **8** (6), 401–417, CrossRef.

Benjamin, S., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev*., **144**, 1669–1694, CrossRef.

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci*., **66**, 603–626, CrossRef.

____, and Coauthors, 2017: Stochastic parameterization: Toward a new view of weather and climate models. *Bull. Amer. Meteor. Soc*., **98**, 565–588, CrossRef.

Bowler, N. E., A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKRB: Upgrades to the MOGERPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc*., **135**, 767–776, CrossRef.

Brown, J. D., D.-J. Seo, and J. Du, 2012: Verification of precipitation forecasts from NCEP's Short Range Ensemble Forecast (SREF) system with reference to ensemble streamflow prediction using lumped hydrologic models. *J. Hydrometeorol*., **13** (3), 808–836, CrossRef.

____, L. Wu, M. He, S. Regonda, H. Lee, and D-J Seo, 2014: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification. *J. Hydrology*, **519**, 2869–288, CrossRef.

Buizza, R., M. Milleer, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc*., **125**, 2887–2908, CrossRef.

Bytheway, J. L., M. Hughes, K. Mahoney, and R. Cifelli, 2020: On the uncertainty of high-resolution hourly quantitative precipitation estimates in California. J. Hydrometeor., 21, 865–879, CrossRef.

____, ____, R. Cifelli, K. Mahoney, and J. M. English, 2022: Demonstrating a probabilistic quantitative precipitation estimate for evaluating precipitation forecasts in complex terrain. *Wea. Forecasting*, **37**, 45–64, CrossRef.

Cannon, F., F. M. Ralph, A. M. Wilson, and D. P. Lettenmaier, 2017: GPM satellite radar measurements of precipitation and freezing level in atmospheric rivers: Comparison with ground-based radars and reanalyses. *Journal of Geophysical Research: Atmospheres*, **122**, 12,747–12,764, CrossRef.

____, J. M. Cordeira, C. W. Hecht, J. R. Norris, A. Michaelis, R. Demirdjian, and F. M. Ralph, 2020: GPM Satellite Radar Observations of Precipitation Mechanisms in Atmospheric Rivers. *Mon. Wea. Rev*., **148**, 1449–1463, CrossRef.

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, CrossRef.

____, and Coauthors, 2022: Advanced quantitative precipitation information: Improving monitoring and forecasts of precipitation, streamflow, and coastal flooding in the San Francisco Bay area, under review.

Corringham, T. W., F. M. Ralph, A. Gershunov, D. R. Cayan, and C. A. Talbot, 2019: Atmospheric rivers drive flood damages in the western United States. *Science Advances,* **5** (12), eaax4631, CrossRef.

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, CrossRef.

DeFlorio, M. J., D. E. Waliser, B. Guan, D. A. Lavers, F. M. Ralph, and F. Vitart, 2018: Global assessment of atmospheric river prediction skill. *Journal of Hydrometeorology*, **19** (2), 409–426.

DeHaan, L. L., A. C. Martin, R. R.Weihs, L. Delle Monache, and F. M. Ralph, 2021: Object-based verification of atmospheric river predictions in the Northeast Pacific. *Weather and Forecasting*, **36** (4), pp.1575–1587.

Demargne, J., J. D. Brown, Y. Liu, D.-J. Seo, L. Wu, Z. Toth, and Y. Zhu, 2010: Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmos. Sci. Lett*., **11** (2), 114–122, CrossRef.

____, and Coauthors, 2014: The science of NOAA's Operational Hydrologic Ensemble Forecast Service. *Bull. Am. Meteorol. Soc*., **95** (1), 79–98, CrossRef.

Dettinger, M., 2011: Climate change, atmospheric rivers and floods in California—A multimodel analysis of storm frequency and magnitude changes. *J. Amer. Water Resour. Assoc*., **47**, 514–523, CrossRef.

____, 2013: Atmospheric rivers as drought busters on the U.S. West Coast. *J. Hydrometeor*., **14**, 1721–1732.

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-Atmospheric Mesoscale Prediction System, *Wea. Forecasting*, **36**, 2275-2288, CrossRef.

Dowell, D. C. and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part 1: Motivation and system description. *Wea. And Forecasting*, **37**, 1371–1395, CrossRef.

Du, J., and Coauthors, 2009: NCEP Short-Range Ensemble Forecast (SREF) system upgrade in 2009. Extended Abstracts, *19th Conf. on Numerical Weather Prediction and 23rd Conf. on Weather Analysis and Forecasting*, Omaha, NE, Amer. Meteor. Soc., 4A.4. [Available online at http://ams.confex.com/ams/23WAF19NWP/techprogram/paper_153264.htm.]

English, J. M., D. D. Turner, T. I. Alcott, W. R. Moninger, J. L. Bytheway, R. Cifelli, and M. Marquis, 2021: Evaluating operational and experimental HRRR model forecasts of atmospheric river events in California. *Weather and Forecasting*, **36**, 1925–1944, CrossRef.

Gershunov, A., and Coauthors, 2019: Precipitation regime change in Western North America: the role of Atmospheric Rivers. *Sci. Rep*., **9**, 9944, CrossRef.

Gimeno, L., R. Nieto, M. Vázquez, and D. A. Lavers, 2014: Atmospheric rivers: A mini-review. Front. *Earth Sci*., **2**, 2.1– 2.6, CrossRef.

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, CrossRef.

Gourley, J. J., Y. Hong, Z. L. 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, CrossRef.

Grim, J. A., J. O. Pinto, T. Blitz, K. Stone, and D. C. Dowell, 2022: Biases in the prediction of convective storm characteristics with a convection allowing ensemble. *Wea. Forecasting*, **37**, 65–83, CrossRef.

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting*, **17**, 192–205, CrossRef.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev*., **125**, 1312–1327, CrossRef.

____, G.T. Bates, J. S.Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Am. Meteorol. Soc*., **94**, 1553–1556, CrossRef.

Herman, G. R. and R. S. Schumacher, 2018: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Monthly Weather Review*, **146**, 1571-1600, CrossRef.

Huang, X., D. L. Swain, D. B. Walton, S. Stevenson, and A. D. Hall, 2020: Simulating and evaluating atmospheric river-induced precipitation extremes along the U.S. Pacific Coast: Case studies from 1980–2017. *Journal of Geophysical Research: Atmospheres*, **125** (4), e2019JD031554, CrossRef.

James, E. P. and coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part II: Forecast performance. *Wea. Forecasting*, **37**, 1397–1417, CrossRef.

Jeworrek, J., G. West, and R. Stull, 2021: WRF precipitation performance and predictability for systematically varied parameterizations over complex terrain. *Wea. Forecasting*, **36**, 893–913, CrossRef.

Juang, H.-M. and M. Kanamitsu, 1994: The NMC nested regional spectral model. *Mon. Wea. Rev*., **122**, 3–26, CrossRef.

____, S.-Y. Hong, and M. Kanamitsu, 1997: The NCEP regional spectral model: An update. *Bull. Amer. Meteor. Soc*., **78**, 2125–2144, CrossRef.

Kalina, E. A., I. Jankov, T. Alcott, J. Olson, J. Beck, J. Berner, D. Dowell, and C. Alexander, 2021: A progress report on the development of the High-Resolution Rapid Refresh ensemble. *Wea. Forecasting*, **36**, 791–804, CrossRef.

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

Lavers, D. A., and Coauthors, 2020: Forecast errors and uncertainties in atmospheric rivers. *Weather and Forecasting*, **35** (4), 1447–1458, CrossRef.

Lewis, W. R., W. J. Steenburgh, T. I. Alcott, and J. J. Rutz, 2017: GEFS precipitation forecasts and the implications of statistical downscaling over the western United States. *Wea. Forecasting*, **32**, 1007–1028, CrossRef.

Lin, X. and A.Y. Hou, 2012: Estimation of rain intensity spectra over the continental United States using ground radar–gauge measurements. *J. Climate*, **25**, 1901-1915, CrossRef.

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

Lundquist, J., M. Hughes, E. Gutmann, and S. Kapnick, 2019: Our skill in modeling mountain rain and snow is bypassing the skill of our observational networks. *Bull. Amer. Meteor. Soc*., **100**, 2473–2490, CrossRef.

Martin, A., F. M. Ralph, R. Demirdjian, L. DeHaan, R. Weihs, J. Helly, D. Reynolds, and S. Iacobellis, 2018: Evaluation of atmospheric river predictions by the WRF Model using aircraft and regional mesonet observations of orographic precipitation and its forcing. *Journal of Hydrometeorology*, **19** (7), 1097–1113, CrossRef.

Mathiesen, P. and J. Kleissl, 2011: Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. *Solar Energy*, **85**, 967–977, CrossRef.

McCollor, D., and R. Stull, 2008: Hydrometeorological short-range ensemble forecasts in complex terrain. Part I: Meteorological evaluation. *Wea. Forecasting*, **23**, 533–556, CrossRef.

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, CrossRef.

Olson, J. B., J. S. Kenyon, W. A. Angevine, J. M. Brown, M. Pagowski, and K. Suselj, 2019: A description of the MYNN-EDMF scheme and the coupling to other components in WRF-ARW. *NOAA Tech. Memo OAR GSD-61*, 37 pp., CrossRef.

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climate prediction models. *Quart. J. Roy. Meteor. Soc*., **127**, 279–304, CrossRef.

Peel, S. and L. J. Wilson, 2008: A Diagnostic Verification of the precipitation forecasts produced by the Canadian Ensemble Prediction System. *Wea. Forecasting*, **23**, 596–616, CrossRef.

Ralph, F. M., P. J. Neiman, and G. A. Wick, 2004: Satellite and CALJET aircraft observations of atmospheric rivers over the eastern North Pacific Ocean during the winter of 1997/98. *Monthly Weather Review*, **132** (7), 1721–1745. CrossRef.

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

Roberts, N. M. and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev*., **136**, 78-97, CrossRef.

Rodwell, M. J., 2006: Comparing and combining deterministic and ensemble forecasts: how to predict rainfall occurrence better. *ECMWF Newsletter*, **106**, 17–23, CrossRef.

Sanchez, C., K. D. Williams, and M. Collins, 2015: Improved stochastic physics schemes for global weather and climate models. *Quart. J. Roy. Meteor. Soc*., **142**, 147–159, CrossRef.

Seo, D.-J., J. Demargne, L. Wu, Y. Liu, J. D. Brown, S. Regonda, and H. Lee, 2010: Hydrologic ensemble prediction for risk-based water resources management and hazard mitigation. In: *4th Federal Interagency Hydrologic Modeling Conference*, Las Vegas, NV, June 27–July 1, 2010, CrossRef.

Skamarock, W. C., and Coauthors, 2019: A Description of the Advanced Research WRF Version 4. *NCAR Tech. Note NCAR/TN-556+STR*, 145 pp., CrossRef.

Smalley, M., T. L'Ecuyer, M. Lebsock, and J. Haynes, 2014: A Comparison of Precipitation Occurrence from the NCEP Stage IV QPE Product and the CloudSat Cloud Profiling Radar. *J. Hydrometeor*., **15**, 444–458, CrossRef.

Stone, R. E., C. A. Reynolds, J. D. Doyle, R. H. Langland, N. L. Baker, D. A. Lavers, and F. M. Ralph, 2020: Atmospheric river reconnaissance observation impact in the Navy Global Forecast System. *Monthly Weather Review*, **148** (2), 763–782, CrossRef.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability*, Reading, United Kingdom, ECMWF, 1–25, CrossRef.

Vokoun, M. and M. Hanel, 2018: Comparing ALADIN-CZ and ALADIN-LAEF precipitation forecasts for hydrological modelling in the Czech Republic. Advances in Meteorology, 2018, Article ID 5368438, 14 pps., CrossRef.

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, CrossRef.

Wilks, D. S., 1995: Statistical Methods in the Atmospheric Sciences. *International Geophysics Series*, **Vol. 59**, Academic Press, 407 pp.

____, 2019: Indices of rank histogram flatness and their sampling properties. *Mon. Wea. Rev*., **147**, 763–769, CrossRef.

Wu., W., D. Kitzmiller, and S. Wu, 2012: Evaluation of radar precipitation estimates from the National Mosaic and Multisensor Quantitative Precipitation Estimation System and the WSR-88D Precipitation Processing System over the Conterminous United States", *J. Hydrometeorology*, **13**, 1080-1093, CrossRef.

Yuan, H., S. L. Mullen, X. Gao, S. Sorooshian, J. Du, and H. H. Juang, 2005: Verification of probabilistic quantitative precipitation forecasts over the Southwest United States during winter 2002/03 by the RSM Ensemble System. *Mon. Wea. Rev*., **133**, 279–294, CrossRef.

Yuan, H., J. A. McGinley, P. J. Schultz, C. J. Anderson, and C. Lu, 2008: Short-range precipitation forecasts from time-lagged multimodel ensembles during the HMT-West-2006 campaign. *J. Hydrometeor*., **9**, 477–491, CrossRef.

Zhang, H. and Z. Pu, 2010: "Beating the Uncertainties: Ensemble Forecasting and Ensemble-Based Data Assimilation in Modern Numerical Weather Prediction", *Advances in Meteorology*, Vol. 2010, Article ID 432160, CrossRef.

Zhao, P., Q.J. Wang, W. Wu, and Q. Yang, 2020: Which precipitation forecasts to use? Deterministic versus coarser-resolution ensemble NWP models. *Quart. J. Roy. Meteor. Soc*., 2021, **147**, 900–913, CrossRef.

Zhu, Y., and R. E. Newell, 1998: A proposed algorithm for moisture fluxes from atmospheric rivers. *Mon. Wea. Rev.,* **126**, 725–735, CrossRef.