Evaluation of Atmospheric River Predictions by the WRF Model Using Aircraft and Regional Mesonet Observations of Orographic Precipitation and Its Forcing

ANDREW MARTIN,[a] F. MARTIN RALPH,[a] REUBEN DEMIRDJIAN,[a] LAUREL DEHAAN,[b] RACHEL WEIHS,[a] JOHN HELLY,[b,c] DAVID REYNOLDS,[d] AND SAM IACOBELLIS[b]

[a] *Center for Western Weather and Water Extremes, Climate Atmospheric Science and Physical Oceanography Division, Scripps Institution of Oceanography, La Jolla, California*
[b] *Climate Atmospheric Science and Physical Oceanography Division, Scripps Institution of Oceanography, La Jolla, California*
[c] *San Diego Supercomputer Center, University of California, San Diego, La Jolla, California*
[d] *Cooperative Institute for Research in Environmental Sciences, Boulder, Colorado*

(Manuscript received 31 May 2017, in final form 24 January 2018)

ABSTRACT

Accurate forecasts of precipitation during landfalling atmospheric rivers (ARs) are critical because ARs play a large role in water supply and flooding for many regions. In this study, we have used hundreds of observations to verify global and regional model forecasts of atmospheric rivers making landfall in Northern California and offshore in the midlatitude northeast Pacific Ocean. We have characterized forecast error and the predictability limit in AR water vapor transport, static stability, onshore precipitation, and standard atmospheric fields. Analysis is also presented that apportions the role of orographic forcing and precipitation response in driving errors in forecast precipitation after AR landfall. It is found that the global model and the higher-resolution regional model reach their predictability limit in forecasting the atmospheric state during ARs at similar lead times, and both present similar and important errors in low-level water vapor flux, moist-static stability, and precipitation. However, the relative contribution of forcing and response to the incurred precipitation error is very different in the two models. It can be demonstrated using the analysis presented herein that improving water vapor transport accuracy can significantly reduce regional model precipitation errors during ARs, while the same cannot be demonstrated for the global model.

1. Introduction

Atmospheric rivers (ARs) play a vital role in delivering rain and snow to western North America (Ralph et al. 2004; Leung and Qian 2009; Guan et al. 2010; Ralph et al. 2010; Dettinger et al. 2011; Neiman et al. 2013). In some regions, as much as 50% of annual precipitation falls on days when an AR is present (Rutz et al. 2014). ARs play an important role in regional recovery from drought (Dettinger 2013) and have been linked to major flooding events in western North America and elsewhere (Ralph et al. 2006; Neiman et al. 2008, 2011; Lavers et al. 2011; Moore et al. 2012). Despite the importance of ARs to water supply and flood risk, the precipitation resulting from ARs remains poorly forecast (Ralph et al. 2010; Junker et al. 2009; Wick et al. 2013b; Lavers et al. 2016). The processes that cause AR precipitation include cloud microphysics (submeter scales), surface moisture fluxes (from submeter to mesoscales), orographic uplift by terrain (from a few kilometers to synoptic scales), and fluid dynamical processes. The latter can operate on the mesoscale (e.g., narrow cold frontal rainbands; Browning and Roberts 1996), synoptic, and global scales. Regardless of the numerical weather prediction (NWP) model one chooses to create a precipitation forecast, some of these scales will not be resolved.

Global numerical weather prediction (GNWP) models, those that can explicitly simulate the largest weather scales on Earth, can also explicitly resolve the scales of mesoscale fluid dynamic features, but do not explicitly resolve the smallest orographic uplift, surface flux, or cloud microphysics scales. To explicitly resolve finer spatial scales, the weather prediction community has traditionally employed regional numerical weather prediction (RNWP) models to dynamically downscale GNWP forecasts [e.g., the National Oceanographic and Atmospheric Administration (NOAA) North American

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-17-0098.s1.

Corresponding author: Andrew Martin, mc@ucsd.edu

DOI: 10.1175/JHM-D-17-0098.1

and High-Resolution Rapid Refresh models (Weygandt et al. 2009), the Deutscher Wetterdienst Consortium for Small-Scale Modeling (Baldauf et al. 2011)]. The RNWP models explicitly resolve finer scales without the aid of empirical parameterization, but most operational weather prediction systems that rely upon RNWP are still unable to explicitly resolve the finest scales (cloud microphysics, surface fluxes, small-scale terrain variability) important to landfalling AR precipitation. In the near future, numerical techniques that allow consistent solutions of the fluid dynamics equations across computational grids with multiple or adjustable resolutions promise to "unify" global and very fine scales in numerical weather prediction [e.g., Model Prediction Across Scales (Skamarock et al. 2012), NOAA's Next Generation Global Prediction System (Michalakes et al. 2015)]. Even with this advance, NWP models will retain an effective limit to explicit spatial resolution that will lie somewhere between scales of topographic variability and fine turbulence scales. We note that while several authors have measured NWP model error for ARs (Junker et al. 2009; Ralph et al. 2010; Kim et al. 2013; Wick et al. 2013b; Swain et al. 2015; Lavers et al. 2016), the predictability limit—the time beyond which it is no longer possible to predict the state of a system, given knowledge of current and past states, with a desired level of accuracy (American Meteorological Society 2017)—for NWP forecasts of AR and their related atmospheric fields has not been established. Additionally, it is not known whether RNWP may improve upon the predictability limit.

Ralph et al. (2013a, hereafter R13) presented a useful way to apportion the impacts of scale-dependent processes on precipitation. The authors used a long record of observations at an atmospheric river observatory (ARO; White et al. 2009, 2013) to demonstrate that the amount of orographic precipitation is linearly related to the bulk upslope flux (BUF) of atmospheric water vapor content by the horizontal wind in a lower tropospheric layer (see their Fig. 5). It has been demonstrated that BUF is also a skillful predictor of instantaneous precipitation rate (Neiman et al. 2002, 2009). Other authors have similarly found that orographic precipitation amount is related to the rate of water vapor flux directed normal to the terrain (Alpert 1986; Barros and Lettenmaier 1994; Sinclair 1994; Colle 2004; Smith and Barstad 2004; Barstad and Smith 2005). These studies examined many regions other than northern coastal California; thus, the importance of low-level moisture flux and the linear relationship found by R13 is applicable to midlatitude orographic precipitation in general.

R13 and Neiman et al. (2009) convincingly demonstrated that the orographic precipitation response to the forcing (vapor transport at approximately low-level jet height) upon a mountain range applied by an AR is linear to first order. If the predictability limit and predictive skill during an AR can be improved, will this lead to more accurate precipitation prediction? An important step in demonstrating this for any modeling system will be to demonstrate that the modeled orographic precipitation response is correct. This precipitation response is somewhat scale dependent, though discussion in Neiman et al. (2002, 2009), White et al. (2009), Smith et al. (2010), and Kingsmill et al. (2013) demonstrates that the horizon between the forcing scale and the response scale is not identical. Nonetheless, a large part of the simulated forcing response relies upon high-resolution terrain, and we thus expect this to improve with higher-resolution modeling systems.

In the current study, we intend to meet the following goals by analyzing forecasts made by two separate models, a GNWP [the Global Forecast System reforecasts (GFSRe; Hamill et al. 2013)] and an RNWP [the Weather Research and Forecast (WRF) Model (Skamarock 2008)]:

1) Characterize the predictability limit of atmospheric state and structure near and within ARs in the midlatitude northeast Pacific Ocean (approximately between latitudes 25° and 45°N). To do so, we will examine the dependence of errors on forecast lead times that are as great as 7 days.
2) Measure errors in the simulated orographic forcing–response relationship during landfalling ARs.
3) Comment on the suitability of RNWP for forecasting precipitation during ARs and use the lessons learned from goals 1 and 2 to suggest model improvements that are most likely to improve precipitation forecasts during landfalling ARs.

We note that all analyses will be supported by observations (direct verification); the effort to address goal 1 represents the most comprehensive direct model verification for AR forecasts by RNWP yet undertaken, and the effort in addressing goal 2 represents a unique method that we believe can be applied to a wide range of numerical weather prediction systems. We also note that evaluating errors incurred by subgrid-scale parameterizations is outside the scope of the current work. All verification analyses are performed with forecasts and observations that are valid over the northeast Pacific or Northern California. Therefore, not all results may generally apply to other regions of the globe affected by ARs. We will make note in the text where results apply beyond the region investigated.

The remainder of this manuscript will be organized as follows. The GFSRe, WRF, and verification datasets used will be described in section 2. Analysis methods, including the method of apportioning precipitation

error by scale, will be described in section 3. Section 4 will present the results of the verification analyses, and section 5 will summarize results and suggest improvements in NWP that may lead to better precipitation forecasts during landfalling ARs.

2. Data and forecast models

a. Atmospheric river observatory

The California Department of Water Resources (DWR) operates an ARO comprising two individual stations, Bodega Bay (BBY) and Cazadero (CZC) in California, as part of the Enhanced Flood Response and Emergency Preparedness (EFREP) network (White et al. 2009, 2013). The BBY station is situated on the coast at sea level and is designed to monitor horizontal vapor flux, integrated water vapor (IWV), and horizontal winds in the atmospheric river low-level jet (LLJ) as it impinges upon the orographically productive coastal mountain ranges. The CZC station is located north of Bodega Bay at the top of a prominent ridge. The CZC station reports precipitation, vertical S-band radar reflectivity, and precipitation drop size distributions during AR conditions. By adapting the techniques of R13, we will use the couplet of stations to investigate orographic forcing (hereafter "forcing") near the coastal edge of the Russian River Watershed (RRW) through bulk upslope flux measured at BBY and orographic precipitation response (hereafter "response") via accumulated precipitation measured at CZC.

BUF is calculated following the methods of Neiman et al. (2002, 2009), in which the controlling layer (800–1200 m MSL) winds are multiplied by the local IWV and projected onto the mountain orthogonal direction. Controlling layer wind is calculated from the 449-MHz wind profiler at BBY. IWV is calculated via radio occultation from the BBY GPS Trimble receiver. Rainfall accumulation at CZC is measured by a tipping-bucket rain gauge and reported to 0.1-mm precision. Hourly, quality-controlled BUF and accumulated rainfall are available from the NOAA Earth System Research Laboratory via anonymous ftp server.

b. GPS-enabled soundings

Airborne dropsondes from the CalWater 2 early start (CWES) and CalWater 2015 (CW2) intensive observing periods (Ralph et al. 2016) are used extensively in our analyses of forecast skill and predictability limit. Each sounding occurred during a midlatitude northeast Pacific AR transect performed by CalWater aircraft. An example, with two transects used in this study, is shown in Fig. 1. Transects were executed to maintain a flight path perpendicular to the direction of troposphere integrated water vapor flux, and 179 sondes from 15 transects taken during 10 separate AR flights are used in this study (Table 1). Herein, an AR is only considered for analysis if maximum integrated vapor transport (IVT; Cordeira et al. 2013) exceeds $500\,\mathrm{kg\,m^{-1}\,s^{-1}}$ by direct observation or by ARO proxy (see section 2c). This threshold will herein be referred to as the "moderate AR" threshold. It reflects the minimum IVT typically leading to significant precipitation upon landfall by a northeast Pacific AR and has emerged through discussions during the first International Conference on Atmospheric Rivers (Ralph et al. 2017). We require a dropsonde transect to cross the full width of the AR core. Herein, the AR core boundaries are defined by the isopleth where IVT exceeds $500\,\mathrm{kg\,m^{-1}\,s^{-1}}$. If the transect includes dropsondes that recorded IVT $\geq 500\,\mathrm{kg\,m^{-1}\,s^{-1}}$ and at least one poleward and one equatorward dropsonde that recorded IVT $< 500\,\mathrm{kg\,m^{-1}\,s^{-1}}$, then the transect is considered to have sampled the full AR core. Last, transect terminal dropsondes must be greater than 100 km from the WRF lateral boundaries to reduce the chance that imprecise advection closure or the boundary damping layer impact the forecast soundings. Soundings were used to estimate AR environment variables: 500-hPa geopotential height z^{500}, IVT, IVW, and 925-hPa equivalent potential temperature θ_e^{925}. Soundings were also used to estimate AR core variables: moist Brunt–Väisälä frequency N_m, which is a measure of static stability that accounts for the effect of moisture, and the discrete (layer) IVT $\mathrm{dIVT} = -(1/g)\int_{p_1}^{p_2}\mathbf{u}q_v\,dp$, where $\mathbf{u}$ is the horizontal wind vector, q_v is the water vapor mixing ratio, and the integral is performed by requiring $p_2 = p_1 - 25$ hPa and iterating p_1 every 25 hPa from 1000 to 325 hPa.

c. Forecast periods

We simulated two groups of ARs occurring between December 2014 and March 2016. These simulations became the forecast datasets that were used to evaluate the GFSRe and WRF predictability limits and to measure errors in the simulated orographic forcing–response relationship for each model. The first group, OCN (for oceanic AR events), included 15 moderate ARs in the midlatitude northeast Pacific Ocean for which GPS-enabled dropsondes were available during a CalWater transect (see Table 1). The OCN airborne observations represent a rich survey of the vertical and transport-normal structures of oceanic AR. The OCN cases were therefore used to investigate the ability of GFSRe and WRF to accurately simulate the transport-normal structure of orographic forcing (e.g., vertical distribution of vapor transport and moist static stability) at forecast lead times up to 7 days.

Many of the ARs in the OCN cases did not cause significant precipitation over land, so a second group of forecast periods [LND (for landfalling AR events);

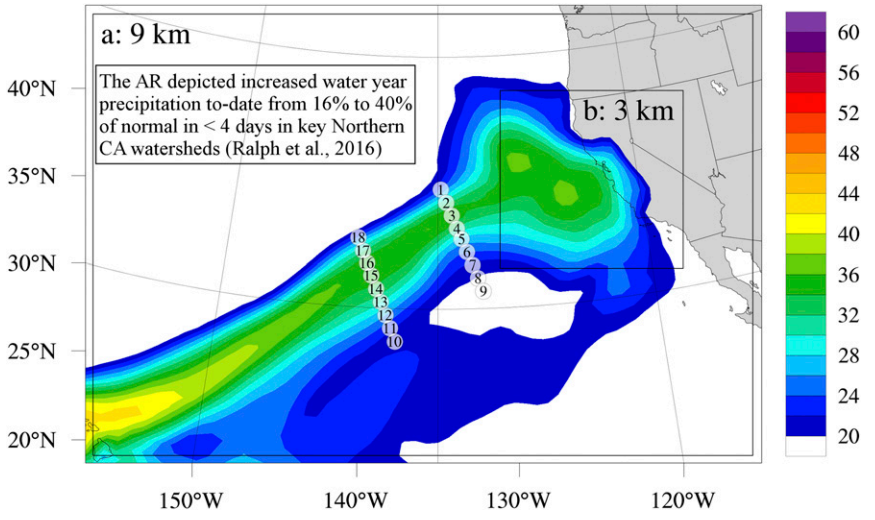

FIG. 1. ERA-Interim reanalysis IWV for 1200 UTC 8 Feb 2014. Overlaid are the dropsondes from two aircraft transects during CalWater early start campaign IOP 2. Individual dropsonde locations are depicted by white circles and numbered in chronological order. These correspond to OCN cases CWES 2 and CWES 3 from Table 1. The text box summarizes some impacts of this AR described in Ralph et al. (2016). Black boxes "a" and "b" correspond to domain boundaries for WRF 9 km and WRF 3 km, respectively.

see Table 2] was chosen to investigate GFSRe and WRF quantitative precipitation forecast (QPF) skill and apportionment of QPF error among scales during moderate ARs impacting the RRW. LND AR cases were chosen using the following criteria:

1) The ARO must have recorded AR conditions following R13 for 24 or more hours.
2) During AR conditions at the ARO, IVT must have exceeded 500 $kg\,m^{-1}\,s^{-1}$, or BUF must have exceeded 300 $mm\,m\,s^{-1}$. The latter has been found to be a proxy for moderate AR conditions.
3) The 10 strongest ARs by storm-integrated BUF that occurred between 1 December 2014 and 31 March 2016 and met criteria 1 and 2 were grouped into the LND case list.

Event start and end for LND cases were declared based upon hourly AR conditions at the ARO following R13. Event duration varies from 24 to 54 h, with a median length of 32 h (Table 2). Storm-total (ST) quantitative precipitation estimation (QPE) within the RRW during LND cases is estimated by the NCEP Stage IV 4-km gridded product (Lin and Mitchell 2005).

TABLE 1. OCN cases simulated for this study. More information regarding IOPs and aircraft can be found in Ralph et al. (2016).

IOP	Aircraft	Transect start	Transect end
CWES 1	NOAA G-IV	2021 UTC 7 Feb 2014	2208 UTC 7 Feb 2014
CWES 2	NOAA G-IV	2050 UTC 8 Feb 2014	2146 UTC 8 Feb 2014
CWES 2	NOAA G-IV	2243 UTC 8 Feb 2014	2338 UTC 8 Feb 2014
CWES 3	NOAA G-IV	1903 UTC 11 Feb 2014	2124 UTC 11 Feb 2014
CWES 4	NOAA G-IV	1734 UTC 12 Feb 2014	1903 UTC 12 Feb 2014
CWES 5	NOAA G-IV	1833 UTC 13 Feb 2014	2058 UTC 13 Feb 2014
CW2 1	NOAA G-IV	2114 UTC 15 Jan 2015	2246 UTC 15 Jan 2015
CW2 1	NOAA G-IV	2307 UTC 15 Jan 2015	0020 UTC 16 Jan 2015
CW2 2	NOAA G-IV	2245 UTC 17 Jan 2015	0030 UTC 18 Jan 2015
CW2 2	NOAA G-IV	0130 UTC 18 Jan 2015	0307 UTC 18 Jan 2015
CW2 4	NOAA G-IV	2004 UTC 24 Jan 2015	2052 UTC 24 Jan 2015
CW2 4	NOAA G-IV	2126 UTC 24 Jan 2015	2303 UTC 24 Jan 2015
CW2 6	NOAA G-IV	2045 UTC 6 Feb 2015	2146 UTC 6 Feb 2015
CW2 6	NOAA G-IV	2159 UTC 6 Feb 2015	2259 UTC 6 Feb 2015
CW2 6	NOAA G-IV	2312 UTC 6 Feb 2015	0004 UTC 7 Feb 2015

TABLE 2. LND cases simulated in this study.

Case	Start at ARO	Duration (h)	NWP valid start
1	1500 UTC 10 Dec 2014	32	1200 UTC 10 Dec 2014
2	0400 UTC 6 Feb 2015	27	0000 UTC 6 Feb 2015
3	0900 UTC8 Feb 2015	25	0600 UTC 8 Feb 2015
4	1300 UTC 9 Dec 2015	26	1200 UTC 9 Dec 2015
5	1400 UTC 20 Dec 2015	47	1200 UTC 20 Dec 2015
6	0400 UTC 17 Jan 2016	25	0000 UTC 17 Jan 2016
7	1700 UTC 28 Jan 2016	32	1800 UTC 28 Jan 2016
8	2200 UTC 5 Mar 2016	33	1800 UTC 5 Mar 2016
9	0800 UTC 9 Mar 2016	42	0600 UTC 9 Mar 2016
10	1500 UTC 12 Mar 2016	37	1200 UTC 12 Mar 2016

d. GEFS reforecasts

The National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) model serves as the GNWP for the tests presented herein. We required that all forecast periods used the same deterministic model. To satisfy this requirement, we acquired control member forecasts from the NOAA Global Ensemble Reforecast Dataset (Hamill et al. 2013). GFSRe forecasts are initialized once per day at 0000 UTC and run to 192 h lead time. GFSRe serves as the WRF parent model in this study. The GFSRe dataset is produced with GEFS version 9.0.1, run at approximately 40-km native resolution. Since 15 January 2015, NOAA has run an updated GFS deterministic forecast with a native resolution of approximately 13 km. Because the majority of the OCN case skill calculations require forecasts initialized before this date, we did not choose the higher-resolution deterministic GFS for the primary GNWP in this study. We have run parallel simulations where possible using the deterministic GFS to verify that key results do not change significantly given the higher-resolution GNWP. The analyses created from these parallel simulations can be found in the online supplemental material. Methods of generating WRF from the higher-resolution GNWP (GFS 0.25; see supplemental material) and methods of postanalysis are identical to those presented herein.

e. WRF

The open-source WRF-ARW model (Skamarock 2008) is used in this study. We configured WRF with two domains utilizing horizontal resolutions of 9 and 3 km (hereafter WRF 9 km and WRF 3 km, respectively). Both WRF 9 km and WRF 3 km are configured with 60 vertical levels with compressed spacing near the 925- and 300-hPa levels in a U.S. Standard Atmosphere sounding. Static land surface information for WRF simulations is generated from the USGS land-use database (Wang et al. 2012) at a resolution of 5 (2) arc-min for WRF 9 km (3 km). WRF domains and parameterized physics options are listed in Table 3. The domains, vertical spacing, and nesting ratio were chosen based on sensitivity tests using the dropsondes from the OCN case list to measure forecast performance in IVT. Parameterized physics options were chosen according to author experience and common practice in other WRF NWP forecast efforts. We stress that the WRF parameterized physics used herein have not been objectively optimized.

The WRF 9 km domain has a much larger Earth-relative footprint (Fig. 1, box ''a'') and utilizes interpolated forecasts from GFSRe as boundary conditions. Initial conditions for both WRF domains are interpolated from the GFSRe analysis to the WRF 9 km and WRF 3 km grids using the WRF preprocessor (Wang et al. 2012). The WRF 3 km domain Earth-relative footprint (Fig. 1, box ''b'') covers most of the state of California and portions of western Nevada and southern Oregon. The RRW, where precipitation is verified in this study, lies near the center of the 3-km domain. The WRF domain configurations described here are also used to create operational forecasts at the Center for Western Weather and Water Extremes (CW3E; http://cw3e.ucsd.edu/). The CW3E operational model, named West-WRF, has the primary goal of predicting extreme precipitation events (especially those associated with ARs) that are key to water supply and flooding in the region (Dettinger et al. 2011; Ralph and Dettinger 2012; R13). As they are further developed, West-WRF operational forecasts will be oriented to the special requirements posed by western U.S. extreme precipitation. These requirements were summarized recently in a study carried out in support of the Western States Water Council's request for a vision for future observational needs for extreme precipitation monitoring, prediction, and climate trend detection (Ralph et al. 2014). This summary built upon more than a dozen reports from various agencies and science groups over the previous few years and on experience

TABLE 3. Domain attributes and parameterized physics options for WRF configurations in this study.

Option	Outermost domain	Nest domain
dx	9 km	3 km
nx; *ny*; *nz*	484; 324; 60	403; 391; 60
Time step	Adaptive, ~45 s	Adaptive, ~15 s
Cumulus	Grell 3D	—
Land surface	Noah	Noah
Cloud microphysics	Thompson New	Thompson New
Planetary boundary layer	Yonsei University	Yonsei University
Surface layer	Monin–Obukhov	Monin–Obukhov
Shortwave radiation	GSFC	GSFC
Longwave radiation	RRTM	RRTM
Topographic wind	No	Yes

in developing California's unique observing network (White et al. 2013) and from NOAA's Hydrometeorology Testbed (Ralph et al. 2013b). In addition to operational forecast goals, West-WRF is designed as a platform from which to evaluate the sources of forecast error and their relationship to physical processes.

3. Methods

a. The verification matrix procedure

The predictability limit (goal 1 in the introduction) has been estimated from a number of forecasts with systematically varying lead time. These forecasts are generated using the verification matrix procedure, described thusly. To estimate the skill for a single event at n lead times, one needs n forecasts that verify at the time of event t_v, each generated at a unique lead time t_i. If this is to be done for m events with l unique observations, one can evaluate the skill at $t_i = 1, \ldots, n$ by generating a set of at most $m \times n$ forecasts. In practice, $N_f < m \times n$ forecasts are started, each from a unique initial time t_0. Forecasts are grouped together according to $t_i = t_v - t_0$. The estimate of forecast skill at each lead time $i = 1, \ldots, n$ will then be estimated from a sample of $m \times l$ observation–forecast pairs. Verification matrices herein are generated by nominal lead times of 24–168 h with an increment of 24. Events are either the period of AR conditions at the ARO or the median time of dropsonde release in an OCN case AR transect. These times do not fall on 0000 UTC. For this reason, we were required to bin our lead times in order to generate a full verification matrix with a nominal time resolution of 24 h. For 24-h resolution, we create a matrix of forecasts sorted into seven lead-time bins. However, for events containing a single observation [such as the orographic forcing and response (section 3b)], we combine every other lead-time bin in order to increase the sample size. The result is a matrix of forecasts sorted into three lead-time bins. Table 4 lists the nominal lead times and their bin boundaries for the seven and three lead-time matrices used in this study.

b. Estimates of predictive skill

For AR environment variables, we define the predictive skill of each model's forecast as the Brier skill score (BSS; Winterfeldt et al. 2011) using GFSRe climatology (1991–2015, December–March) as the reference forecast. The BSS is equivalent to measuring the fractional reduction in error variance by the more accurate forecast. The predictability limit in any AR environment variable is considered to be reached when BSS ≤ 0. GFSRe climatology is not available at full vertical resolution, and therefore it is not feasible to estimate the model climatology for the AR core variables. Instead, the predictability limit is crossed for dIVT and N_m^2 when the Student's t test is significant at $p < 0.05$ when comparing forecasts of the OCN case core variables to their observed counterparts. This is equivalent to choosing the following desired level of forecast accuracy: the probability that OCN case forecasts are chosen from the same distribution as the OCN case observations remains greater than 5%.

Qualitative and quantitative methods are used to assess accuracy in QPF for the models considered. We compute histograms from each model's QPF for each lead time in the three lead-time verification bins and compare to histograms computed from the NCEP Stage IV ST QPE. The population is drawn from all events and all Stage IV grid

TABLE 4. Verification matrices (see section 3a) used in this study.

Bin	7-bin boundaries (h)	3-bin boundaries (h)
1	$12 \leq t_i \leq 35$	$12 \leq t_i \leq 59$
2	$36 \leq t_i \leq 59$	$60 \leq t_i \leq 107$
3	$60 \leq t_i \leq 83$	$108 \leq t_i \leq 155$
4	$84 \leq t_i \leq 107$	—
5	$108 \leq t_i \leq 131$	—
6	$132 \leq t_i \leq 155$	—
7	$156 \leq t_i \leq 179$	—

points within the RRW. For quantitative measures, we calculate the mean, standard deviation, and root-mean-square of the quantity QPF − QPE at all NCEP Stage IV grid points and for all models and lead times as appropriate. Bilinear interpolation is used to transform QPF on the WRF 3 km and GFSRe grids to the NCEP Stage IV grid.

We also estimate the normalized error of GFSRe and WRF 3 km in reproducing the observed storm-total bulk upslope flux of moisture (ST BUF) and precipitation P_r relationship at the ARO (goal 2 in the introduction). We follow R13 in defining a forcing–response relationship wherein the forcing is ST BUF and the response is storm-total precipitation (ST P_r). The total accuracy in simulating this relationship will be measured by the multifactor orographic forcing–response error:

$$e_{xy} = e_x + e_y = E\left[\frac{(x - x_o)^2}{V[x_o]} + \frac{(y - y_o)^2}{V[y_o]}\right], \quad (1)$$

where x and y are the forcing (ST BUF) and ST P_r, respectively. The subscript o refers to the observed values of the given quantity, and $E[\ ]$ and $V[\ ]$ refer to the expectation and variance operators, respectively.

c. *Linearization of the forcing and response relationships, reduction in error*

We also assess whether reducing error in the forcing or reducing error in the simulated response, independent of forcing, leads to a larger reduction of error in the forecast ST P_r.

Let $e_y = E[(y - y_o)^2]/V[y_o]$ represent the normalized mean-square error in the forecast ST P_r compared to that measured by the ARO. We can also define

$$e_y \sim \frac{E[F(x) - y_o^2]}{V[y_o]}, \quad (2)$$

where $F(x) \sim f(x)$ is the linearized approximation of the response function described in R13. We can estimate $F(x)$ from each model and from observations by least squares approximation. An example is shown in Fig. 2a. The figure shows the ST BUF and ST P_r measured at the ARO during 171 rain events (gray dots). Here, $F(x)$ is approximated by applying a least squares fit the individual data points. The result is the red line in the figure. Also shown in Fig. 2a are observed (orange circle), GFSRe (orange asterisk), and WRF 9 km ST P_r and ST BUF for a single LND case evaluated in this study. The line segments A and B on the figure graphically represent the nondimensional distance that e_{xy} is designed to measure. Let the quantity $F_o(x)$ represent the "perfect response" approximation. It is the ST P_r that results from applying the linearized local response function derived from a least-squared fit of the observations to the forecast forcing. The fractional reduction in error in ST P_r by perfect response approximation is then

$$\delta e_{\text{ypr}} = 1 - \frac{E[F_o(x) - y_o^2]}{V[y_o]e_y}. \quad (3)$$

Conversely, $F(x_o)$ represents the "perfect forcing" approximation. The fractional reduction in error in ST P_r by perfect forcing is

$$\delta e_{\text{ypf}} = 1 - \frac{E[F(x_o) - y_o^2]}{V[y_o]e_y}. \quad (4)$$

This process is visualized in Fig. 2b. The black line represents the least squares approximation of a hypothetical set of ST P_r observations to a set of hypothetical ST BUF observations. In this hypothetical case, all observations lie on the line, such that the normalized error e_y of the least squares fit is 0. The slope of the line represents $F(x)_o$. The blue line and blue diamonds are a set of hypothetical model forecasts of ST P_r and ST BUF. While the model underpredicts ST BUF, its response function (slope) is higher than the observed. This hypothetical set of forecasts has $e_y = 1.548$, suggesting that the mean error in forecast ST P_r is 154% of the ST P_r variance. The remaining lines in Fig. 2b represent a perfect forcing correction (green upward triangles) and perfect response correction (purple downward triangles) applied to the hypothetical model, respectively. Each has a different impact (δe_{ypf}, δe_{ypr}) on the resulting normalized error in predicted ST P_r, shown in the inset of the figure. For this hypothetical case, correcting the model response function lowers prediction error, while correcting the forcing greatly increases the prediction error, because the flawed model response function in the hypothetical example causes much too great precipitation response given realistic forcing. The function $F(x)_o$ is estimated from the sample of historical observations at the ARO that reside within the domain (range) of ST BUF (ST P_r) defined by the LND cases. This corresponds to a sample size of 52. The function $F(x)$ is estimated for WRF and GFSRe by each LND case forecast binned by the three lead-time verification bins. This procedure yielded a sample size of 20 for each WRF and GFSRe.

d. *Sonde data processing*

Each dropsonde is processed by vertical smoothing onto isobaric surfaces every 25 hPa from 1000 to 300 hPa. The diagnostic variables calculated include IVT, moist Brunt–Väisälä frequency N_m^2 following Ralph et al. (2005), and equivalent potential temperature θ_e as approximated in Stull (2012). The dropsondes report

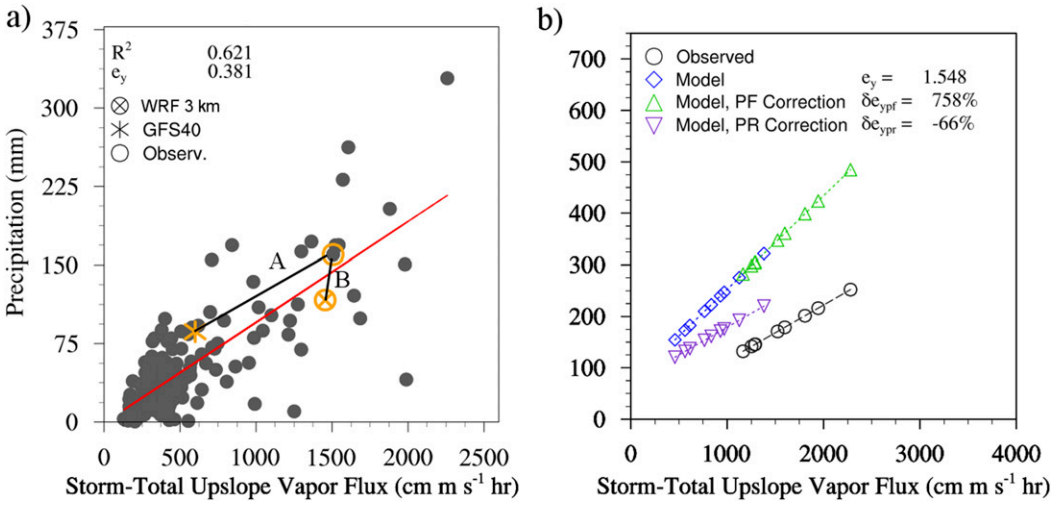

FIG. 2. (a) ST P_r and ST BUF at the ARO for 171 historical rain events (dark gray dots) and the linear trend line resulting from a least squares fit (red line). The correlation coefficient R^2 and e_y that result from the least squares fit are provided. Also provided is one example from the LND case list of the observed (orange circle), GFSRe forecast (orange asterisk) and WRF 3 km forecast (orange circle and cross). The segments A and B represent the nondimensional distance measured by e_{xy}. (b) Example of a set of observed (black circles) ST P_r and ST BUF and the linear relationship found from a least squares fit (black line). The blue diamonds (blue line) show a set of hypothetical model predictions of ST P_r and ST BUF and its linear fit. Parameter e_y from the hypothetical model predictions is provided in the inset. Green (purple) lines and upward (downward) triangles represent the perfect forcing (response) linear correction to the hypothetical model prediction. Parameters δe_{ypf} and δe_{ypr} resulting from the perfect forcing and response corrections are also provided in the inset.

an observation approximately every 1 s; however, we assign a static observation time for each sonde that corresponds to the mean time of its transect.

Two of the 15 total transect observing periods required unique processing. Those are transect 1 on 7 February 2014 and transect 4 on 11 February 2014. Both of these transects are a composite of two spatially offset flight tracks.

e. Transect compositing procedure

To derive AR-normal composite cross sections, it was necessary to align the dynamical features within each AR transect. First, each individual dropsonde is linearly interpolated to a gridded AR-normal transect with a resolution of 50 km. Second, a common center among the transects is defined as the dropsonde with maximum IVT and the endpoints defined by most poleward and equatorward sondes. Finally, a composite of the transects is created by taking the arithmetic mean of the interpolated transects.

f. NWP forecast to observation interpolation

1) SPATIAL INTERPOLATION TO SONDES AND TEMPORAL INTERPOLATION

Forecasts from GFSRe and WRF 9 km have been spatially interpolated from their native grids to the dropsonde Earth-relative location using a bilinear method. There are two sources of temporal uncertainty in our methods. First, no attempt has been made to temporally interpolate NWP output to sonde report time. To create composite dropsonde transects, we must assume stationarity in the AR up to the longest time between transect start and end. This time is 2 h and 27 min, or very near the 3-h output interval for each model. Second, the GFSRe duty cycle is 24 h, but the valid time of any given observation may occur anywhere in the diurnal cycle. The sources of temporal uncertainty above lead primarily to random temporal imprecision in model to observation matching, though we cannot rule out that these sources of temporal imprecision will accumulate to nonzero residual. Because their temporal offset is likely to be much larger than the time a dropsonde takes to profile the atmosphere below flight level, no effort is made to account for the sonde drift in the forecast interpolation. The location used is the mean of the individual sonde latitude and longitude reports.

2) ATMOSPHERIC RIVER OBSERVATORY

NWP forecast output from all models was bilinearly interpolated to both ARO locations. Model output was also linearly interpolated from the model native coordinate to the effective retrieval heights (m AGL) of the 449-MHz wind profiling radar.

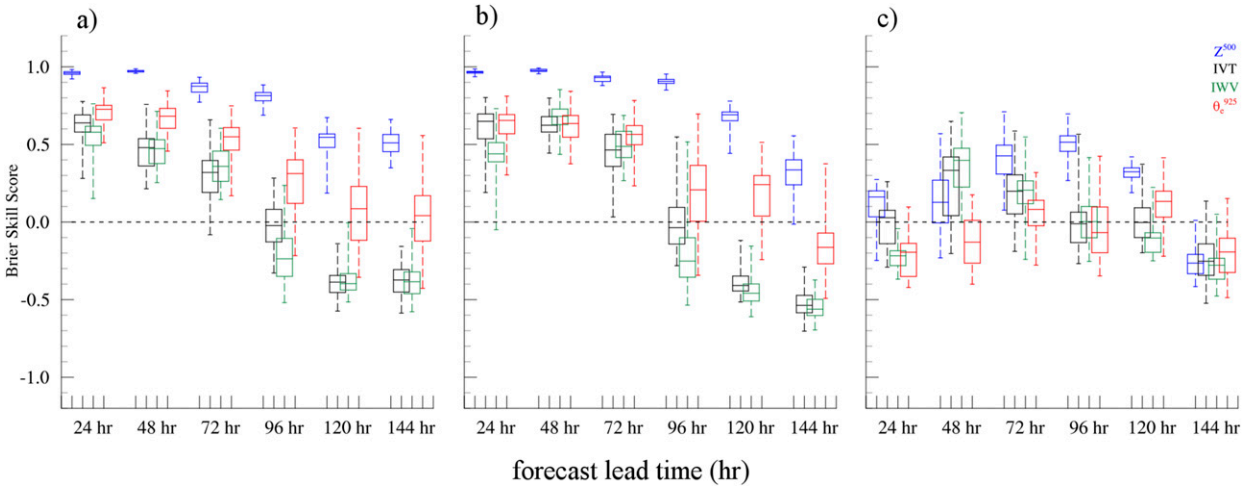

FIG. 3. (a) Value added by GFSRe over GFSRe climatology validated against 145 CalWater 2 dropsondes for the state variables z^{500} (blue), IVT (black), IWV (green), and θ_e^{925} (red). (b) As in (a), but the variable is WRF 9 km value added over GFSRe climatology. (c) As in (b), but the reference forecast is GFSRe.

3) NCEP STAGE IV QPE

Model output was bilinearly interpolated from native grids to the locations of the NCEP Stage IV QPE product. Model output was masked thereafter to exclude any points lying outside the boundaries of the RRW. Watershed boundaries were defined by a georeferenced shapefile created by the USGS. Masking by the shapefile polygon was performed using NCAR Command Language version 6.2 (http://www.ncl.ucar.edu/).

4. Results

a. *Predictability limit in AR state variables*

GFSRe predictive skill for lead times 24–144 h is displayed in Fig. 3a. BSS is estimated for z^{500}, θ_e^{925}, IWV, and IVT as described in section 3b. Only sondes for which IVT $\geq 250\,\mathrm{kg\,m^{-1}\,s^{-1}}$ were retained for the analysis, yielding a sample of 145 dropsondes. Upper (lower) whiskers represent maximum (minimum) BSS, upper (lower) box bounds represent upper (lower) quartile BSS, and the box center line represents median BSS.

For $t_i < 84$ h, GFSRe forecasts of these state variables add value to a climatology forecast. Forecasts of z^{500} have the highest expected skill, and skill remains very high even for longer lead times. IWV and IVT display the least skill and most skill degradation. These variables are those traditionally used to identify and track ARs (Dettinger 2011; Lavers et al. 2012; Wick et al. 2013a; Guan et al. 2013), and thus we consider the GFSRe predictability limit to exist in the window 84 h $\leq t_i \leq$ 107 h. Figure 3b similarly shows predictive skill for WRF 9 km. Behavior is not significantly different for most variables and lead times, and we consider the predictability limit to similarly be met for 84 h $\leq t_i \leq$ 107 h.

The value added by WRF 9 km to GFSRe for the first six lead times considered is displayed in Fig. 3c. From Fig. 3c it is apparent that WRF 9 km adds value to GFSRe for AR state variables between 36 h $\leq t_i \leq$ 83 h. Details such as maximum value added and the range of lead times for which value is added/lost vary by variable, with most value added for z^{500} and least for θ_e^{925}.

We recreated the analysis in Fig. 3 using GFS 0.25 as the GNWP and West-WRF 9 km. This companion analysis is shown in the supplemental material as Fig. S3. When the parent model is GFS 0.25, WRF 9 km adds value for 84 h $\leq t_i$ for forecasts of IWV, IVT, and θ_e^{925}. For z^{500}, the value-added range is delayed to 108 h $\leq t_i$. For some variables, notably IVT, BSS varied considerably with lead time. This may be a result of the smaller number CalWater 2 sonde matching forecasts available for GFS 0.25. We do not declare a predictability limit for the GFS 0.25 forecasts for this reason.

b. *Vertical structures of AR static stability*

We investigated the predictive skill in forecasts of moist static stability by GFSRe and WRF 9 km from the sample of dropsondes for which IVT $\geq 500\,\mathrm{kg\,m^{-1}\,s^{-1}}$. This analysis is shown in Fig. 4 using the normalized mean-square error and bias in forecast moist Brunt–Väisälä frequency N_m^2. The center panel displays a vertical profile of median CalWater dropsonde N_m^2 observations evaluated every 25 hPa from 1000 to 600 hPa. The typical value of N_m^2 reported in Ralph et al. (2005) varies near $1–2 \times 10^{-4}\,\mathrm{s^{-2}}$ from the ocean's surface to

600 hPa. The measurements we report here appear to be slightly higher near the surface but relax to a similar profile above 900 hPa. We attribute this difference to different sonde selection criteria and potentially different AR environments. The remaining panels display WRF 9 km (left side) and GFSRe (right side) normalized mean-square error and normalized bias in N_m^2 for the three lead-time verification bins. Lead time increases from the top of Fig. 4 toward the bottom. Normalized mean-square error is calculated following the formula for e_y, with forecast (observed) N_m^2 standing in for y (y_o), respectively. Normalized bias is calculated as $E[y - y_o]/\sqrt{V[y_o]}$. The mean normalized bias in the GFSRe (WRF 9 km) forecasts is very similar for all lead times, both in maximum/minimum bias and in mean profile. Forecasts by WRF 9 km and GFSRe are both too unstable at levels below 850 hPa. Biases then become positive as model pressure decreases toward 600 hPa. Normalized mean-square errors e_y typically maximize near 925 hPa in both models across all lead times. This is near the mean LLJ level in the composite observational transect (Fig. 5). The exception is e_y for GFSRe for 12 h $\leq t_i \leq$ 59 h. In this profile (upper-right panel), mean-square errors maximize at 2.5 times the observational variance ($\sim 5.5 \times 10^{-4}\,\mathrm{s}^{-4}$) at the upper end of the profile. For all other lead times, GFSRe mean-square error profiles are similar to WRF 9 km, but maximum errors are consistently greater in GFSRe by a fraction of the observational variance. The average simulated AR core is therefore more unstable than it should be near the AR LLJ. If moisture transport is properly simulated, this unstable bias may lead to more precipitation than should be expected by orographic uplift of a moist-neutral layer, though the effect should be similar in both models.

c. *Vertical structures of AR core horizontal vapor transport*

Figure 5a displays the composite AR cross sections of dIVT from the OCN case AR core transects. The methodology for constructing each cross section is discussed in section 3f. The center panel in Fig. 5 displays the observed composite. In this panel, there is a strong local maximum in water vapor flux near 900 hPa located at the analyzed AR core center. This water vapor flux maximum is located just below a weak composite LLJ (isotach composite, not shown). Equivalent potential temperature isotherms in Fig. 5 show the composite AR core straddling a baroclinic zone with temperatures decreasing poleward of the core center. This composite structure is consistent with AR cross sections reported in Cordeira et al. (2013) and Ralph et al. (2016).

The remaining panels in Fig. 5 display the mean dIVT error (forecast minus observed) composite in WRF and GFSRe for the three lead-time verification bins. Overlaid in each is the composite forecast θ_e. It is apparent that these errors in water vapor flux are spatially heterogeneous and that errors are more likely to be positive above 700 hPa and more likely to be negative near the mean LLJ position for both models. Composites were made for each individual variable contributing to water vapor flux (wind speed and water vapor mixing ratio, not shown). It was found that the errors in each are spatially correlated with each other and with dIVT. Model levels at pressures less than 700 hPa in each model composite are too moist and wind speeds are too fast, while the LLJ region in each model composite is too dry and wind speeds are too slow. Several authors (Thorpe and Clough 1991; Dudhia 1993; Lafore et al. 1994; Wakimoto and Murphey 2008) have observed a subgeostrophic polar jet in conjunction with a supergeostrophic LLJ in midlatitude cyclones containing strong cold fronts. The results reported here are consistent with the hypothesis that the model is too geostrophic and may be inadequately representing the ageostrophic winds. The upper troposphere positive bias in dIVT is significant at the $p < 0.05$ level at short lead times only. Negative biases in dIVT near the mean LLJ position do not become significant at the $p < 0.05$ level until the later lead times. In the absence of model climatology for LLJ dIVT, the significance of the t test can be interpreted as indicating the predictability limit has been exceeded for these regions. Negative biases in dIVT are strongest in the LLJ, a feature found to be critically important to driving heavy orographic precipitation (Browning et al. 1974; Bader and Roach 1977; Neiman et al. 2002; Smith et al. 2010), and are quite similar in spatial extent and magnitude for both models.

d. *Deterministic QPF skill during landfalling AR*

Value histograms for WRF 3 km and GFSRe ST QPF at NCEP Stage IV grid points within the RRW for the three lead-time verification bins are shown in Fig. 6. Figure 6 also displays the LND case ST QPE histogram. ST QPE from the cases investigated had median and upper (lower) quartile values of 61 and 97 (41) mm. For short lead times (Fig. 6a), the WRF 3 km QPF histogram most closely resembles the QPE histogram. The inset tables in Fig. 6 display mean QPF − QPE (Bias), the standard deviation of QPF − QPE (σ), and the root-mean-square error (RMSE) QPF − QPE for both WRF 3 km and GFSRe. Both WRF 3 km and GFSRe produce low-biased ST QPF across lead times. For $t_i \leq$ 59 h, WRF 3 km accumulation bias is 5% greater in magnitude relative to median QPE than for GFSRe. For other lead times, bias becomes similar.

For $t_i \leq$ 108 h, WRF 3 km produces a smaller range in accumulation error (σ is 10% smaller relative to median

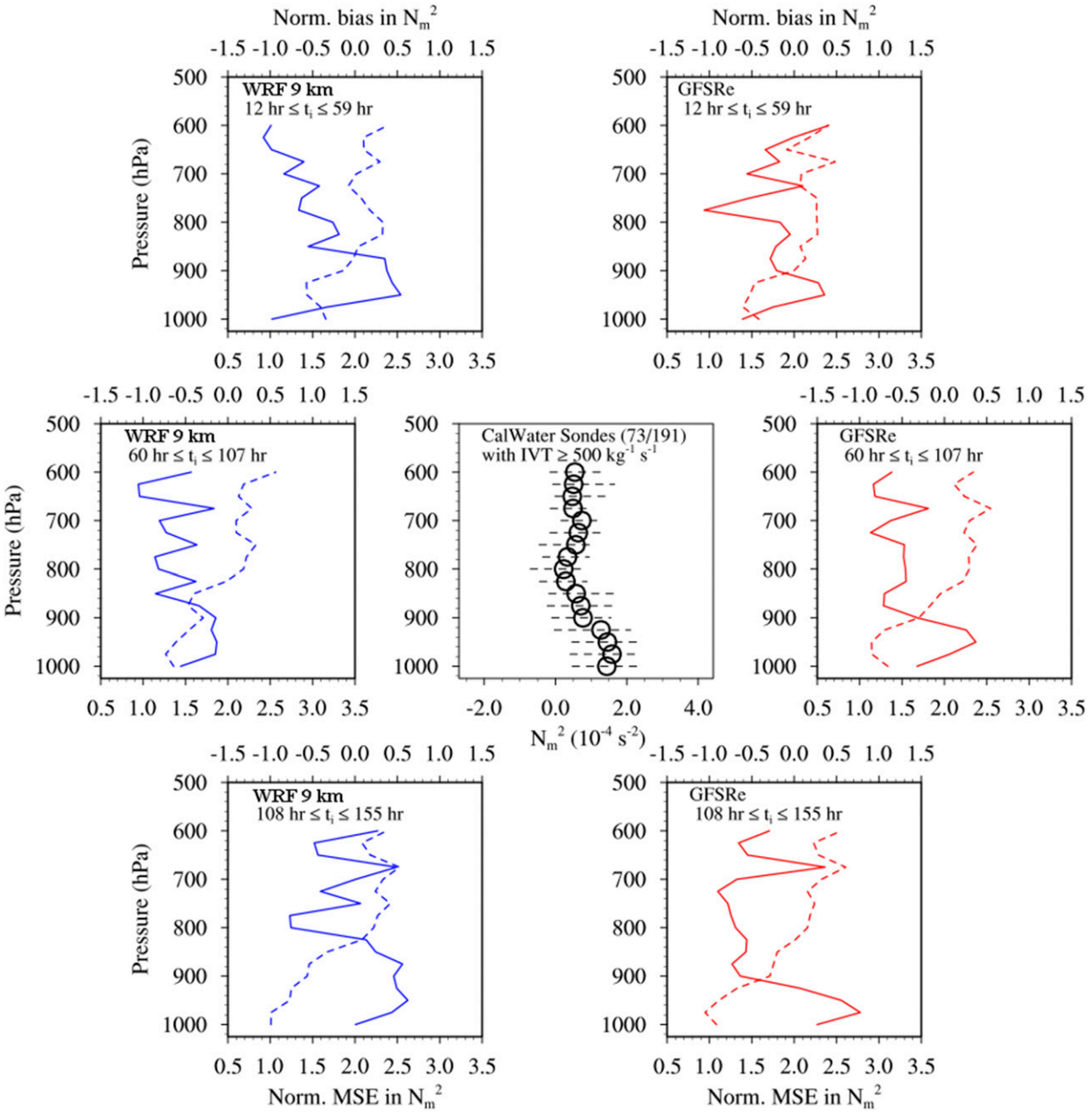

FIG. 4. (top) Normalized mean-square error [Eq. (2)] in N_m^2 (solid, bottom axis) and normalized bias $E[y - y_o]/\sqrt{V[y_o]}$ in N_m^2 (dashed, top axis) for 12 h $\leq t_i \leq$ 59 h. The left panel displays mean error in the WRF 9 km (blue) interpolated soundings, and the right panel displays the same quantity for GFSRe (red). (middle) As in the top row, but for 60 h $\leq t_i \leq$ 107 h in the left and right panels. The center panel displays the observed median N_m^2 (10^{-4} s^{-2}, black circles) from all CalWater sondes satisfying the "AR core" criterion (section 3b). The interquartile range in N_m^2 is shown by dashed lines. (bottom) As in the top row, but for 108 h $\leq t_i \leq$ 155 h.

QPE). For $t_i \leq$ 108 h, WRF 3 km RMSE is 5%–7% less than GFSRe as a fraction of median QPE, while for $t_i >$ 108 h, GFSRe displays an advantage that is similar in magnitude. WRF 3 km produces total precipitation greater than 140 mm (near the upper 10% value for QPE) with nonzero frequency. GFSRe, however, does not produce these upper 10% values. This last advantage WRF displays is likely related to higher spatial resolution.

Analysis from Fig. 6 does not definitively answer whether WRF 3 km or GFSRe QPF is more accurate for RRW landfalling AR. Neither WRF 3 km nor GFSRe distinguished itself in a large or consistent manner over the lead times considered. Thus far, we have seen that

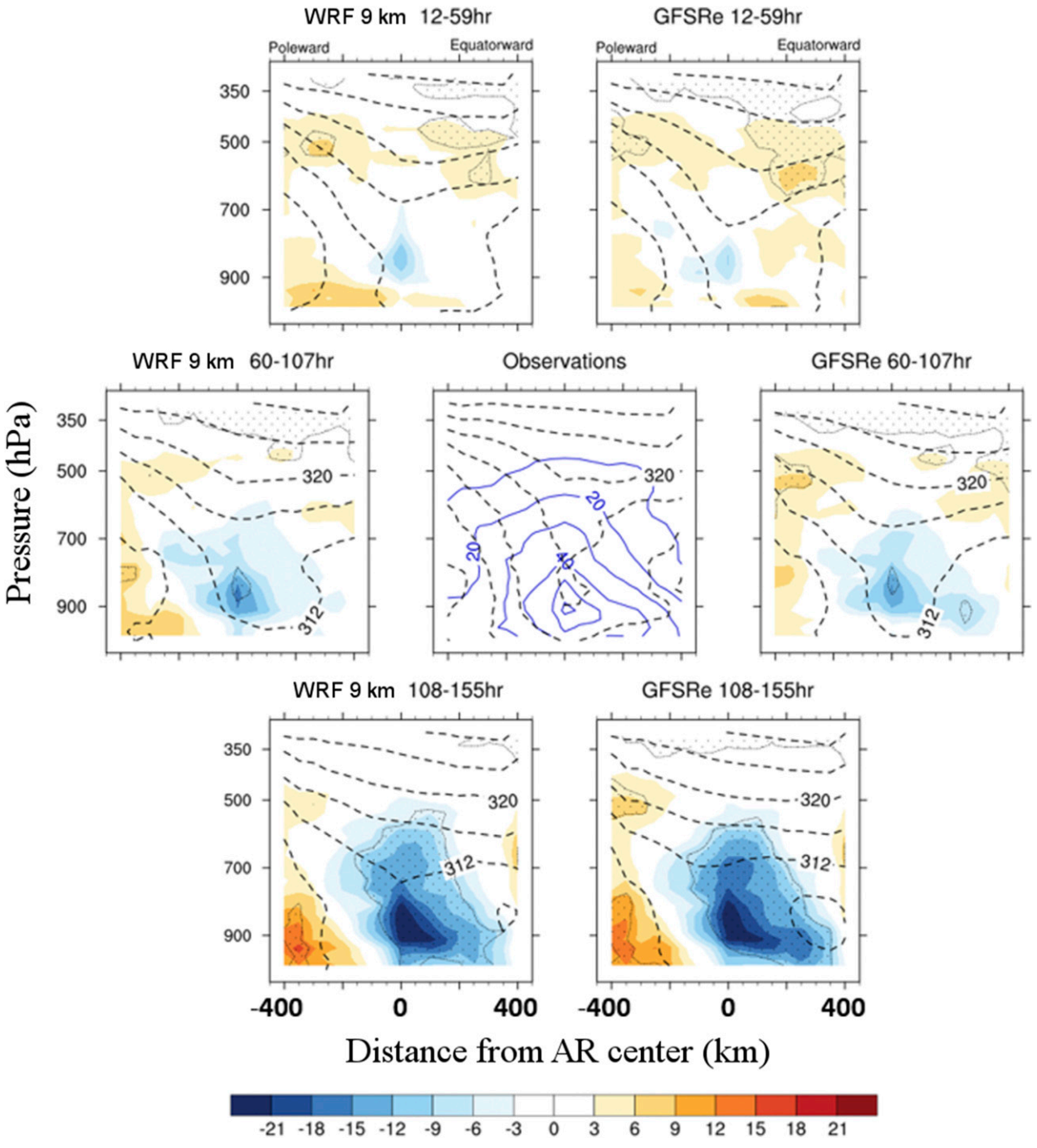

FIG. 5. (top) OCN case mean AR normal–vertical transect of dIVT error (model minus observation; kg m^{-1} s^{-1}; color fill), and θ_e^{925} (K; black dashed lines) for 12 h $\leq t_i \leq$ 59 h. The left panel displays mean error in the WRF 9 km transects, and the right panel displays same quantity for GFSRe. Stippling indicates significance at $p < 0.05$ according to a Student's t test. (middle) As in the top row, but for 60 h $\leq t_i \leq$ 107 h in the left and right panels. The center panel displays the observed ensemble mean dIVT (kg m^{-1} s^{-1}; blue solid lines) and θ_e^{925} (K; black dashed lines). (bottom) As in the top row, but for 108 h $\leq t_i \leq$ 155 h.

the WRF and GFSRe systems display minor strengths relative to each other in AR state, structure, and RRW QPF, but on balance perform with similar accuracy. However, we cannot simply assume that the QPF performance result follows from the predictive skill in dIVT and static stability. Each model may arrive at its simulated precipitation uninformed by the correct orographic forcing–response relationship. If so, we cannot

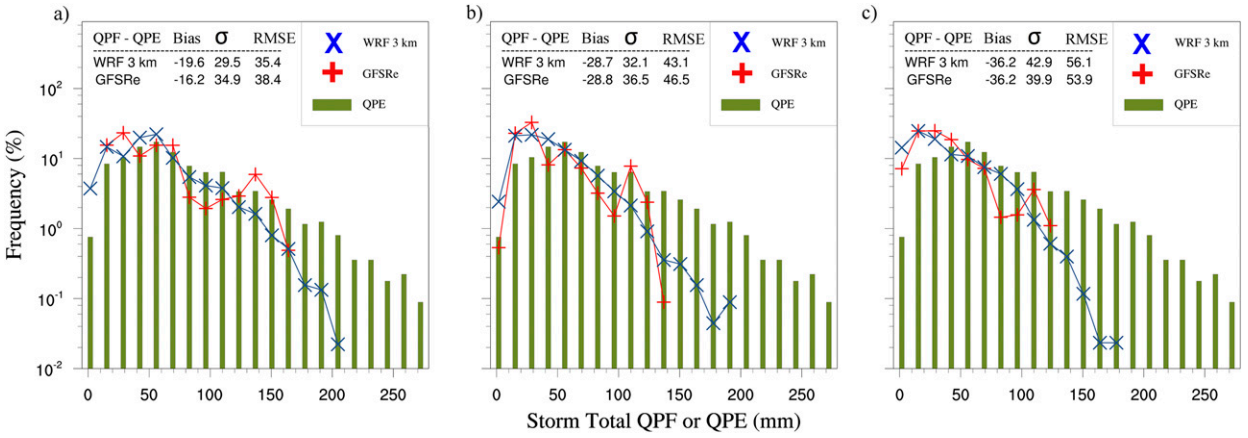

FIG. 6. (a) Value histogram for RRW ST QPE (NCEP Stage IV; green bars), WRF (blue x), and GFSRe (red +). ST QPF calculated from 12 h $\leq t_i \leq$ 59 h. The inset shows WRF and GFSRe mean Bias, σ, and RMSE QPF − QPE from all RRW Stage IV grid points. (b) As in (a), but for 60 h $\leq t_i \leq$ 107 h. (c) As in (a), but for 108 h $\leq t_i \leq$ 155 h.

expect improvements in model forcing accuracy (e.g., improvements in the skill from Figs. 4 and 5) to lead to improvements in precipitation skill.

e. *Relationships between orographic forcing and response*

To adequately address the second goal posed in the introduction, we must investigate each model's ability to accurately reproduce the observed forcing–response relationship during an AR. To this end, we turn to the 10-yr record of bulk upslope vapor flux and mountaintop precipitation collected at the ARO. Figure 7a shows the observed, GFSRe simulated, and WRF 3 km simulated ST BUF–ST P_r relationship for 60 h $\leq t_i \leq$ 107 h e_{xy} for WRF 3 km and GFSRe in the inset table. Parameter e_{xy} is also listed for the three lead-time verification bins in Table 5. Note that WRF 3 km reduces e_{xy} by 37%–80% compared to GFSRe. The reader can qualitatively assess the accuracy of each model in representing the multifactor forcing–response error by matching a triplet of observation, GFSRe, and WRF 3 km symbols for a given color. A unique color is assigned to each LND event forecast or observation; otherwise, colors in Fig. 7a are meaningless. Table 5 additionally presents e_y, the error in storm-total precipitation for both models normalized to the observed variance. WRF 3 km also outperforms GFSRe by 18%–69% in this metric.

From Fig. 7a and Table 5 we conclude that WRF 3 km better predicts the multifactor orographic forcing–response at the ARO. We are also interested in asking the following questions of both models: do errors in ST P_r during ARs arise primarily due to errors in forcing or in the response, and which model may benefit from adjustments to accuracy in either forcing or the response relationship? To investigate this, we present δe_{ypf} and δe_{ypr} (see section 3c) in Table 5.

To interpret the δe_y metrics, we must also compare the model's linearized response relationship (slope; also reported in Table 5) to the observed. Figure 7b graphically displays the linearized response relationship from observations, GFSRe forecasts, and WRF 3 km forecasts at lead times 60 h $\leq t_i \leq$ 107 h. The markers display the least squares derived ST P_r for the given ST BUF. The slope and the correlation coefficient resulting from the least squares fit are also provided in the inset. Note that the WRF 3 km slope is much closer to the observed than is the GFSRe. As we will see, the more accurate response relationship simulated in WRF 3 km will allow both improvements in forcing accuracy or response accuracy to result in improved ST P_r prediction.

WRF δe_{ypf} and δe_{ypr} both become negative for $t_i \geq$ 60 h. At these lead times, both ST BUF and ST P_r are biased low compared to observations (e.g., WRF least squares fit line is low and to the left of observations in Fig. 7b), but the slope of the response relationship is very near that of the observed. This suggests that if WRF forecast ST BUF improves, more accurate WRF ST P_r would result, as reflected by δe_{ypf} for WRF at $t_i \geq$ 60 h in Table 5. WRF δe_{ypr} suggests that ST P_r could also improve by further tuning the precipitation response relationship, though not by as much. For WRF at $t_i \leq$ 60 h, imposing observational ST BUF or response relationship does not improve forecast ST P_r. At these lead times e_y is less than 1.0, suggesting that WRF forecast error is smaller than the observed variance in precipitation and substituting the observed forcing (response) is not effective.

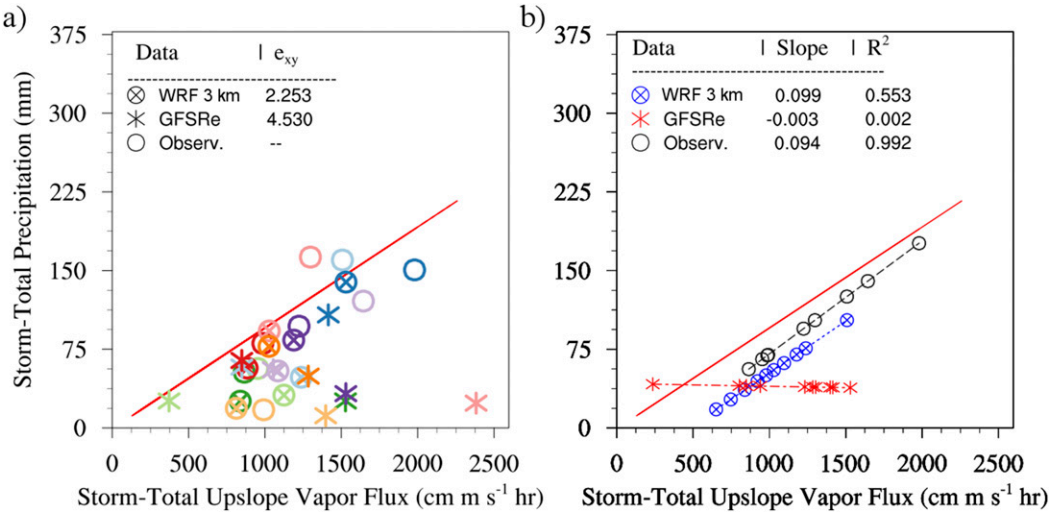

FIG. 7. (a) ST P_r and ST BUF at the ARO for all LND observed (colored circles), GFSRe forecasts (colored asterisk), and WRF forecasts (colored circle and cross) with lead times 60 h $\leq t_i \leq$ 83 h. The inset table displays e_{xy} for each model. The red line is the trend line from all historical ARO events as shown in Fig. 2a. A unique color is assigned to each LND event forecast or observation. The specific color carries no special meaning. (b) As in (a), but black line displays linear fit of LND observed ST P_r to observed ST BUF. Black circles are data points regressed to lie along the best fit line. The red line (asterisks) similarly displays the linear fit derived from GFSRe forecasts and the blue line (circle and cross symbols) similarly displays the linear fit of WRF 9 km forecasts. The inset table displays slope and correlation coefficients of the linear fits.

In contrast, GFSRe δe_{ypf} is nonnegative and GFSRe $e_y > 1.0$ for all t_i. This suggests that the GFSRe response relationship is not accurate enough to translate improved ST BUF into improved ST P_r. Examining Table 5 confirms that the slope of the derived linear relationship for GFSRe is not similar to the observed at any t_i. GFSRe δe_{ypr} suggests that accuracy in ST P_r could be improved if the response relationship is made more accurate, though not to the same degree as for WRF.

We also recreated the analyses in Fig. 7 and Table 5 using GFS 0.25 and the companion WRF 3 km forecasts using GFS 0.25 as initial and boundary conditions. These analyses can be found in Fig. S7 and Table S6, respectively. The broad results from the above section apply when comparing GFS 0.25 to its downscaled WRF forecasts: WRF reduces e_{xy} by as much as 81%. Similarly, substituting the observed forcing or response relationship reduces precipitation error in WRF, but only substituting the observed response relationship reduces precipitation error in GFS 0.25 forecasts. This latter point is likewise because the linearized response relationship for GFS 0.25 renders precipitation insensitive

TABLE 5. Measures of error for ST storm-scale forcing and local-scale response, as well as reduction of error in response by prescribing linearized ARO observed forcing and response to GFSRe and WRF LND case forecasts.

Lead time (h)		$12 \leq t_i \leq 59$	$60 \leq t_i \leq 107$	$108 \leq t_i \leq 155$
Error measure (see sections 3b, 3c)	Model			
e_{xy}	GFSRe	4.287	4.530	7.370
	WRF	0.820	2.253	4.653
e_y	GFSRe	1.544	2.072	2.568
	WRF	0.470	1.295	2.109
δe_{ypf}	GFSRe	18.1%	14.2%	0.0%
	WRF	10.7%	−57.6%	−78.1%
δe_{ypr}	GFSRe	18.9%	−35.0%	−19.3%
	WRF	20.4%	−47.2%	−51.9%
Linear slope (mm s cm^{-1} m^{-1})	Model/obs			
	GFSRe	0.003	−0.003	0.005
	WRF	0.084	0.099	0.091
	Obs	0.094	0.094	0.094

to BUF. The primary difference in the companion analysis presented in the supplemental material is that the linearized response relationship for WRF forced by GFS 0.25 changes from its GFSRe-forced counterpart, from 0.091 to 0.049 mm s cm^{-1} m^{-1}.

5. Discussion

This study is the first to investigate predictability limit of AR state and AR core variables by GNWP and RNWP and to relate deficiencies in those to AR precipitation skill by measuring the error in the simulated orographic forcing–response relationship. It was found that WRF 9 km is capable of adding value to its parent GNWP by means of dynamical downscaling for a subset of medium-range weather prediction time scales. This result suggests that assimilation of observations into the coarse domain at model native scales may additionally improve forecasts created by WRF 9 km.

Generally, WRF 9 km and GFSRe forecasts of AR vertical and transport-normal structures were found to reproduce realistic vapor transport in the LLJ region of ARs at short lead times, but forecasts of LLJ water vapor flux were found to develop significant low bias by $t_i > 108$ h lead times. Forecasts of moist static stability in the ARs were too unstable at lower levels and too stable at pressures below 600 hPa, but these forecast errors were similar for both models, especially at short lead times for all models. The GNWP considered (the Global Forecast System) is capable of producing ARs realistic in structure at sufficiently short lead time. If well constructed, RNWP forecasts downscaled from the same GNWP can as well.

WRF 3 km and GFSRe displayed similar accuracy in predicting RRW storm-total precipitation during the AR considered. As lead time increased, both models produced a significant dry bias compared to the observed accumulated precipitation distribution. This finding may follow from the inability of both models to produce a strong AR LLJ at longer lead times (Fig. 5). The authors note that tuning of the parameterized physics submodels in WRF for the purpose of accurately predicting AR precipitation has not been done and that QPF smoothed to lower resolutions (e.g., the difference in resolution between WRF 3 km and GFSRe) has been shown to result in higher skill scores over complex terrain (Mass et al. 2002).

When predicting storm-total precipitation at a mountaintop well known to be orographically productive (CZC), WRF improves upon GFSRe mean-square error by as much as 69% at short lead times. It appears that this improvement occurs primarily because WRF better reproduces the relationship between orographic forcing (approximated by ST BUF) and response (ST P_r at CZC). This can be seen visually in Figs. 7a and 7b and quantitatively in Table 5. WRF 3 km very accurately reproduces the ST BUF–ST P_r relationship found through observations during the AR cases studied.

It is found that improvement in WRF QPF can be expected through either more accurate forcing (e.g., data assimilation) or response (e.g., parameterized physics tuning). WRF forcing at the ARO was often low-biased, in agreement with the low-bias in LLJ dIVT (Fig. 7). Thus, the consistent underprediction of RRW ST P_r (Fig. 6) is partially caused by storm-scale forcing that is too weak. This cause-and-effect relationship cannot be verified for GFSRe, since ST BUF errors from GFSRe are more randomly distributed (Fig. 7b) and since the local response relationship is not similar to that observed.

The analysis presented herein suggests that improvements to either forcing or orographic precipitation response will straightforwardly lead to more accurate precipitation in WRF or similar RNWP. This is true even at lead times approaching 7 days. Therefore, WRF may be an attractive option to produce skillful QPF for regions in which heavy rain events are dominated by atmospheric rivers, especially given intensive work to develop data assimilation techniques to address the low bias found in LLJ water vapor flux and to develop more accurate parameterizations of key subgrid-scale processes such as surface energy fluxes and cloud microphysics.

Acknowledgments. The U.S. Army Corps of Engineers Award W912HZ-15-SOI-0019, the National Science Foundation XSEDE Award ATM150010, and the California Department of Water Resources provided financial support for this research. The authors would also like to thank the NOAA CalWater airborne science teams, the NOAA Physical Science Division's Hydrometeorological Testbed, and the University of California, Davis, Bodega Marine Laboratory for providing data and facilities. Additionally, the authors thank Scripps Institution of Oceanography for establishing the Center for Western Weather and Water Extremes.

REFERENCES

Alpert, P., 1986: Mesoscale indexing of the distribution of orographic precipitation over high mountains. *J. Climate Appl. Meteor.*, **25**, 532–545, https://doi.org/10.1175/1520-0450(1986)025<0532:MIOTDO>2.0.CO;2.

American Meteorological Society, 2017: Predictability limit. Glossary of Meteorology, http://glossary.ametsoc.org/wiki/Predictability_limit.

Bader, M., and W. Roach, 1977: Orographic rainfall in warm sectors of depressions. *Quart. J. Roy. Meteor. Soc.*, **103**, 269–280, https://doi.org/10.1002/qj.49710343605.

Baldauf, M., A. Seifert, J. Förstner, D. Majewski, M. Raschendorfer, and T. Reinhardt, 2011: Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Mon. Wea. Rev.*, **139**, 3887–3905, https://doi.org/10.1175/MWR-D-10-05013.1.

Barros, A. P., and D. P. Lettenmaier, 1994: Dynamic modeling of orographically induced precipitation. *Rev. Geophys.*, **32**, 265–284, https://doi.org/10.1029/94RG00625.

Barstad, I., and R. B. Smith, 2005: Evaluation of an orographic precipitation model. *J. Hydrometeor.*, **6**, 85–99, https://doi.org/10.1175/JHM-404.1.

Browning, K., and N. Roberts, 1996: Variation of frontal and precipitation structure along a cold front. *Quart. J. Roy. Meteor. Soc.*, **122**, 1845–1872, https://doi.org/10.1002/qj.49712253606.

——, F. Hill, and C. Pardoe, 1974: Structure and mechanism of precipitation and the effect of orography in a wintertime warm sector. *Quart. J. Roy. Meteor. Soc.*, **100**, 309–330, https://doi.org/10.1002/qj.49710042505.

Colle, B. A., 2004: Sensitivity of orographic precipitation to changing ambient conditions and terrain geometries: An idealized modeling perspective. *J. Atmos. Sci.*, **61**, 588–606, https://doi.org/10.1175/1520-0469(2004)061<0588:SOOPTC>2.0.CO;2.

Cordeira, J. M., F. M. Ralph, and B. J. Moore, 2013: The development and evolution of two atmospheric rivers in proximity to western North Pacific tropical cyclones in October 2010. *Mon. Wea. Rev.*, **141**, 4234–4255, https://doi.org/10.1175/MWR-D-13-00019.1.

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

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

——, F. M. Ralph, T. Das, P. J. Neiman, and D. R. Cayan, 2011: Atmospheric rivers, floods and the water resources of California. *Water*, **3**, 445–478, https://doi.org/10.3390/w3020445.

Dudhia, J., 1993: A nonhydrostatic version of the Penn State–NCAR Mesoscale Model: Validation tests and simulation of an Atlantic cyclone and cold front. *Mon. Wea. Rev.*, **121**, 1493–1513, https://doi.org/10.1175/1520-0493(1993)121<1493:ANVOTP>2.0.CO;2.

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

——, ——, ——, ——, and ——, 2013: The 2010/2011 snow season in California's Sierra Nevada: Role of atmospheric rivers and modes of large-scale variability. *Water Resour. Res.*, **49**, 6731–6743, https://doi.org/10.1002/wrcr.20537.

Hamill, T. M., 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. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

Junker, N. W., M. J. Brennan, F. Pereira, M. J. Bodner, and R. H. Grumm, 2009: Assessing the potential for rare precipitation events with standardized anomalies and ensemble guidance at the hydrometeorological prediction center. *Bull. Amer. Meteor. Soc.*, **90**, 445–454, https://doi.org/10.1175/2008BAMS2636.1.

Kim, J., D. E. Waliser, P. J. Neiman, B. Guan, J. M. Ryoo, and G. A. Wick, 2013: Effects of atmospheric river landfalls on the cold season precipitation in California. *Climate Dyn.*, **40**, 465–474, https://doi.org/10.1007/s00382-012-1322-3.

Kingsmill, D. E., P. J. Neiman, B. J. Moore, M. Hughes, S. E. Yuter, and F. M. Ralph, 2013: Kinematic and thermodynamic structures of sierra barrier jets and overrunning atmospheric rivers during a landfalling winter storm in Northern California. *Mon. Wea. Rev.*, **141**, 2015–2036, https://doi.org/10.1175/MWR-D-12-00277.1.

Lafore, J., J. Redelsperger, C. Cailly, and E. Arbogast, 1994: Nonhydrostatic simulation of frontogenesis in a moist atmosphere. Part III: Thermal wind imbalance and rainbands. *J. Atmos. Sci.*, **51**, 3467–3485, https://doi.org/10.1175/1520-0469(1994)051<3467:NSOFIA>2.0.CO;2.

Lavers, D. A., R. P. Allan, E. F. Wood, G. Villarini, D. J. Brayshaw, and A. J. Wade, 2011: Winter floods in Britain are connected to atmospheric rivers. *Geophys. Res. Lett.*, **38**, L23803, https://doi.org/10.1029/2011GL049783.

——, G. Villarini, R. P. Allan, E. F. Wood, and A. J. Wade, 2012: The detection of atmospheric rivers in atmospheric reanalyses and their links to British winter floods and the large-scale climatic circulation. *J. Geophys. Res.*, **117**, D20106, https://doi.org/10.1029/2012JD018027.

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

Leung, L. R., and Y. Qian, 2009: Atmospheric rivers induced heavy precipitation and flooding in the western U.S. simulated by the WRF regional climate model. *Geophys. Res. Lett.*, **36**, L03820, https://doi.org/10.1029/2008GL036445.

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.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? The results of two years of real-time numerical weather prediction over the Pacific Northwest. *Bull. Amer. Meteor. Soc.*, **83**, 407–430, https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Michalakes, J., and Coauthors, 2015: Evaluating performance and scalability of candidate dynamical cores for the next generation global prediction system. *MultiCore 5 Workshop*, Boulder, CO, NCAR, 15 pp., https://www2.cisl.ucar.edu/sites/default/files/Michalakes_Slides.pdf.

Moore, B. J., P. J. Neiman, F. M. Ralph, and F. E. Barthold, 2012: Physical processes associated with heavy flooding rainfall in Nashville, Tennessee, and vicinity during 1–2 May 2010: The role of an atmospheric river and mesoscale convective systems. *Mon. Wea. Rev.*, **140**, 358–378, https://doi.org/10.1175/MWR-D-11-00126.1.

Neiman, P. J., F. M. Ralph, A. White, D. Kingsmill, and P. 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.

——, ——, G. A. Wick, J. D. Lundquist, and M. D. Dettinger, 2008: Meteorological characteristics and overland precipitation impacts of atmospheric rivers affecting the west coast of North America based on eight years of SSM/I satellite observations. *J. Hydrometeor.*, **9**, 22–47, https://doi.org/10.1175/2007JHM855.1.

——, A. B. White, F. M. Ralph, D. J. Gottas, and S. I. Gutman, 2009: A water vapour flux tool for precipitation forecasting.

Proc. Inst. Civ. Eng.: Water Manage., **162**, 83–94, https://doi.org/10.1680/wama.2009.162.2.83.

——, L. J. Schick, F. M. Ralph, M. Hughes, and G. A. Wick, 2011: Flooding in western Washington: The connection to atmospheric rivers. *J. Hydrometeor.*, **12**, 1337–1358, https://doi.org/10.1175/2011JHM1358.1.

——, F. M. Ralph, B. J. Moore, M. Hughes, K. M. Mahoney, J. M. Cordeira, and M. D. Dettinger, 2013: The landfall and inland penetration of a flood-producing atmospheric river in Arizona. Part I: Observed synoptic-scale, orographic, and hydrometeorological characteristics. *J. Hydrometeor.*, **14**, 460–484, https://doi.org/10.1175/JHM-D-12-0101.1.

Ralph, F. M., and M. Dettinger, 2012: Historical and national perspectives on extreme west coast precipitation associated with atmospheric rivers during December 2010. *Bull. Amer. Meteor. Soc.*, **93**, 783–790, https://doi.org/10.1175/BAMS-D-11-00188.1.

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

——, ——, 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.

——, ——, G. A. Wick, S. I. Gutman, M. D. Dettinger, D. R. Cayan, and A. B. White, 2006: Flooding on California's Russian River: Role of atmospheric rivers. *Geophys. Res. Lett.*, **33**, L13801, https://doi.org/10.1029/2006GL026689.

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

——, T. Coleman, P. Neiman, R. Zamora, and M. Dettinger, 2013a: Observed impacts of duration and seasonality of atmospheric-river landfalls on soil moisture and runoff in coastal Northern California. *J. Hydrometeor.*, **14**, 443–459, https://doi.org/10.1175/JHM-D-12-076.1.

——, J. Intrieri, D. Andra Jr., S. Boukabara, and D. Bright, 2013b: The emergence of weather-focused testbeds 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.

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

——, and Coauthors, 2017: Atmospheric rivers emerge as a global science and applications focus. *Bull. Amer. Meteor. Soc.*, **98**, 1969–1973, https://doi.org/10.1175/BAMS-D-16-0262.1.

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

Sinclair, M. R., 1994: A diagnostic model for estimating orographic precipitation. *J. Appl. Meteor.*, **33**, 1163–1175, https://doi.org/10.1175/1520-0450(1994)033<1163:ADMFEO>2.0.CO;2.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

——, J. B. Klemp, M. G. Duda, L. D. Fowler, S.-H. Park, and T. D. Ringler, 2012: A multiscale nonhydrostatic atmospheric model using centroidal Voronoi tesselations and C-grid staggering. *Mon. Wea. Rev.*, **140**, 3090–3105, https://doi.org/10.1175/MWR-D-11-00215.1.

Smith, B. L., S. E. Yuter, P. J. Neiman, and D. Kingsmill, 2010: Water vapor fluxes and orographic precipitation over Northern California associated with a landfalling atmospheric river. *Mon. Wea. Rev.*, **138**, 74–100, https://doi.org/10.1175/2009MWR2939.1.

Smith, R. B., and I. Barstad, 2004: A linear theory of orographic precipitation. *J. Atmos. Sci.*, **61**, 1377–1391, https://doi.org/10.1175/1520-0469(2004)061<1377:ALTOOP>2.0.CO;2.

Stull, R. B., 2012: *An Introduction to Boundary Layer Meteorology*. Springer, 670 pp.

Swain, D. L., B. Lebassi-Habtezion, and N. S. Diffenbaugh, 2015: Evaluation of nonhydrostatic simulations of northeast Pacific atmospheric rivers and comparison to in situ observations. *Mon. Wea. Rev.*, **143**, 3556–3569, https://doi.org/10.1175/MWR-D-15-0079.1.

Thorpe, A. J., and S. A. Clough, 1991: Mesoscale dynamics of cold fronts: Structures described by dropsoundings in FRONTS 87. *Quart. J. Roy. Meteor. Soc.*, **117**, 903–941, https://doi.org/10.1002/qj.49711750103.

Wakimoto, R. M., and H. V. Murphey, 2008: Airborne Doppler radar and sounding analysis of an oceanic cold front. *Mon. Wea. Rev.*, **136**, 1475–1491, https://doi.org/10.1175/2007MWR2241.1.

Wang, W., and Coauthors, 2012: ARW version 3.4 modeling system user's guide. NCAR, 384 pp., http://www2.mmm.ucar.edu/wrf/users/docs/user_guide_V3.4/ARWUsersGuideV3.pdf.

Weygandt, S. S., T. Smirnova, S. Benjamin, K. Brundage, S. Sahm, C. Alexander, and B. Schwartz, 2009: The High Resolution Rapid Refresh (HRRR): An hourly updated convection resolving model utilizing radar reflectivity assimilation from the RUC/RR. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 15A.6, https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154317.htm.

White, A. B., F. Ralph, P. Neiman, D. Gottas, and S. Gutman, 2009: The NOAA coastal atmospheric river observatory. *34th Conf. on Radar Meteorology*, Williamsburg, VA, Amer. Meteor. Soc., 10B.4, https://ams.confex.com/ams/34Radar/techprogram/paper_155601.htm.

——, 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.

Wick, G. A., P. J. Neiman, and F. M. Ralph, 2013a: Description and validation of an automated objective technique for identification and characterization of the integrated water vapor signature of atmospheric rivers. *IEEE Trans. Geosci. Remote Sens.*, **51**, 2166–2176, https://doi.org/10.1109/TGRS.2012.2211024.

——, ——, ——, and T. M. Hamill, 2013b: 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.

Winterfeldt, J., B. Geyer, and R. Weisse, 2011: Using QuikSCAT in the added value assessment of dynamically downscaled wind speed. *Int. J. Climatol.*, **31**, 1028–1039, https://doi.org/10.1002/joc.2105.