⬯The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part II: Forecast Performance

Eric P. James,[a,b] Curtis R. Alexander,[b] David C. Dowell,[b] Stephen S. Weygandt,[b]
Stanley G. Benjamin,[b] Geoffrey S. Manikin,[c] John M. Brown,[b] Joseph B. Olson,[b] Ming Hu,[b]
Tatiana G. Smirnova,[a,b] Terra Ladwig,[b] Jaymes S. Kenyon,[a,b] and David D. Turner[b]

[a] *Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado*
[b] *NOAA/Global Systems Laboratory, Boulder, Colorado*
[c] *NOAA/Environmental Modeling Center, College Park, Maryland*

ABSTRACT: The High-Resolution Rapid Refresh (HRRR) is a convection-allowing implementation of the Advanced Research version of the Weather Research and Forecast (WRF-ARW) Model that covers the conterminous United States and Alaska and runs hourly (for CONUS; every 3 h for Alaska) in real time at the National Centers for Environmental Prediction. The high-resolution forecasts support a variety of user applications including aviation, renewable energy, and prediction of many forms of severe weather. In this second of two articles, forecast performance is documented for a wide variety of forecast variables and across HRRR versions. HRRR performance varies across geographical domain, season, and time of day depending on both prevalence of particular meteorological phenomena and the availability of both conventional and nonconventional observations. Station-based verification of surface weather forecasts (2-m temperature and dewpoint temperature, 10-m winds, visibility, and cloud ceiling) highlights the ability of the HRRR to represent daily planetary boundary layer evolution and the development of convective and stratiform cloud systems, while gridded verification of simulated composite radar reflectivity and quantitative precipitation forecasts reveals HRRR predictive skill for summer and winter precipitation systems. Significant improvements in performance for specific forecast problems are documented for the upgrade versions of the HRRR (HRRRv2, v3, and v4) implemented in 2016, 2018, and 2020, respectively. Development of the HRRR model data assimilation and physics paves the way for future progress with operational convective-scale modeling.

SIGNIFICANCE STATEMENT: NOAA's operational hourly updating convection-allowing model, the High-Resolution Rapid Refresh (HRRR), is a key tool for short-range weather forecasting and situational awareness. Improvements in assimilation of weather observations, as well as in physics parameterizations, has led to improvements in simulated radar reflectivity and quantitative precipitation forecasts since the initial implementation of HRRR in September 2014. Other targeted development has focused on improved representation of the diurnal cycle of the planetary boundary layer, resulting in improved near-surface temperature and humidity forecasts. Additional physics and data assimilation changes have led to improved treatment of the development and erosion of low-level clouds, including subgrid-scale clouds. The final version of HRRR features storm-scale ensemble data assimilation and explicit prediction of wildfire smoke plumes.

1. Introduction

On 30 September 2014, the first version of the High-Resolution Rapid Refresh (HRRR) numerical weather prediction (NWP) system was operationally implemented at the U.S. National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Prediction (NCEP). Building upon

⬯Denotes content that is immediately available upon publication as open access.

Kenyon's current affiliation: NOAA/National Weather Service Forecast Office, Grand Rapids, Michigan.

Corresponding author: Eric P. James, eric.james@noaa.gov

the successful implementation of the hourly updated 13-km Rapid Refresh (RAP) system in 2012 (e.g., Benjamin et al. 2016, hereafter B16), this implementation marked a major milestone in U.S. NWP as the HRRR provided the first operational hourly updated convection-allowing model (CAM) guidance covering the conterminous United States. The HRRR now represents a critical prediction tool for operational forecasters in the United States, and serves as an initial baseline capability for future generations of CAM guidance to build upon. The first article of this two-part series (Dowell et al. 2022, hereafter D22) documents the system design of the HRRR model; in this article, we quantify forecast performance by the HRRR in its various incarnations from version 1 to version 4.

The goal of the HRRR system is to analyze and predict, as accurately as possible, the short-range (0–18 h) evolution of the atmosphere, to support weather-sensitive decisions. The
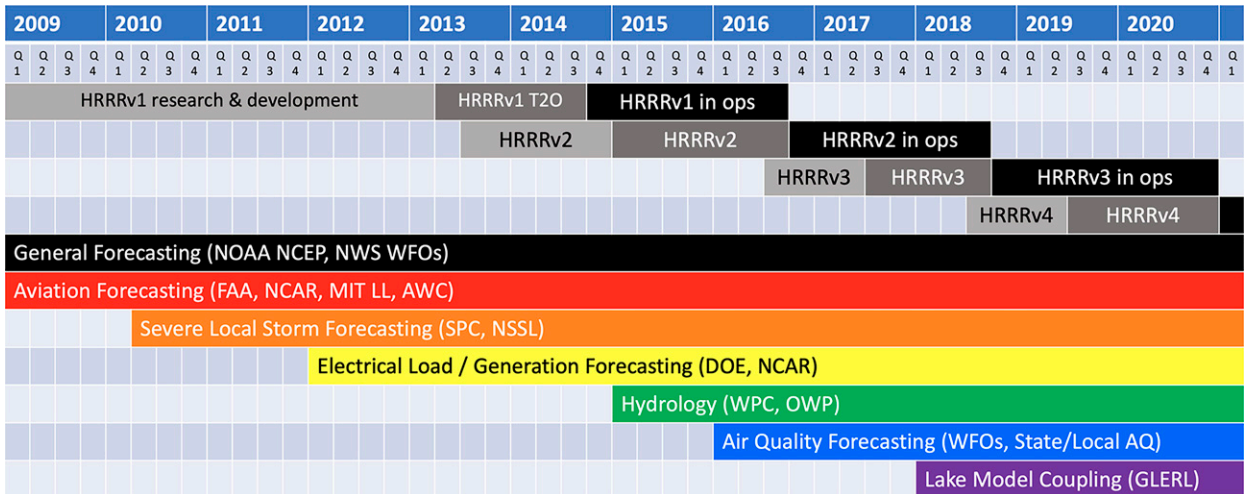
FIG. 1. Timeline of HRRR development and the consideration of new forecast applications for the HRRR's rapidly updating prediction system. Light gray boxes indicate the research and development period for each HRRR version, darker gray boxes indicate the transition to operations, and solid black boxes indicate HRRR versions running in operations. Key collaborators are mentioned for each forecast application area.

relatively fine grid spacing (3 km) in the HRRR allows deep, moist convection to evolve explicitly on the model grid; this permits much more realistic forecasts compared to models with parameterized convection, not just for convective storms but also for other phenomena such as orographic precipitation, mesoscale precipitation bands, and local wind circulations. Since our Earth's atmosphere affects many sectors of society, the HRRR necessarily has many stakeholders from various fields of analysis and forecast application. A timeline of the development of the various incarnations of the HRRR in Fig. 1 presents its emerging new applications over the years. As the HRRR's treatment of atmospheric physics and Earth system processes has increased in complexity from version to version, with the implementation of a host of changes with each new version, new stakeholders and collaborators have begun to use HRRR forecasts for their applications. The interconnected nature of Earth system components often means that improved forecasts for one particular application have positive impacts for other forecast applications, but also increases the number of metrics needed for evaluation of each new version. Accomplishing improvements across all application areas is a challenging task, and development efforts need to be carefully coordinated. Changes are prioritized which target known biases, with additional efforts focused on exploring other potential data assimilation (DA) or physics improvements. User requests for specific additional output, such as increased temporal or vertical resolution, or output variables requiring inline computation, are considered where possible, but are often a lower priority than opportunities for more advanced DA or physics approaches.

A number of researchers have undertaken to evaluate HRRR forecasts for individual phenomena or particular applications. Table 1 summarizes HRRR evaluation studies of which the authors are aware. Phenomena investigated in these articles range from winter precipitation type and amounts (e.g., Ikeda et al. 2013, 2017; Dougherty et al. 2021; English

et al. 2021) to convective storm prediction (e.g., Pinto et al. 2015; Bytheway and Kummerow 2015; Bytheway et al. 2017; Duda and Turner 2021), to detailed evaluations of planetary boundary layer (PBL) structure (e.g., Fovell and Gallagher 2020) and surface fluxes (e.g., Lee et al. 2019). In this paper, we provide a more comprehensive quantification of HRRR forecast performance, with a particular emphasis upon demonstrating the improvements stemming from subsequent versions of the HRRR.

Changes toward HRRR improvements have been designed through a combination of statistical evaluation and ongoing discussions on HRRR behavior with users from NWS and others in the aviation, energy, and severe weather communities. Feedback from forecasters on deficiencies in HRRR forecasts is often tied to individual events, but over time consistent signals emerge which help target improvements for subsequent versions. It is important to note that the HRRR cannot be expected to perfectly capture every event due to the chaotic nature of the atmosphere, and it often exhibits run-to-run variability in timing, spatial extent, or even occurrence for high-impact events, particularly in more uncertain situations. This highlights the need for convection-allowing ensembles [e.g., the NOAA High Resolution Ensemble Forecast (HREF; Jirak et al. 2018) now including HRRRv4] and statistical postprocessing [e.g., the National Center for Atmospheric Research neural network probability forecasts (Sobash et al. 2020) also using HRRRv4] to avoid overreliance on deterministic HRRR predictions.

In this article, evaluation is generally carried out based on real-time results from both the experimental and operational HRRR, but results from controlled retrospective tests are also shown where appropriate. Section 2 describes forecast performance, beginning with verification of simulated radar reflectivity forecasts and quantitative precipitation forecasts (QPF), followed by surface weather forecasts (2-m temperature

TABLE 1. Previously published articles evaluating specific aspects of HRRR performance. "Retro" indicates the data were taken from a set of retrospective simulations for a particular case or cases.

Study	Focus	HRRR version	Region	Period of study
Ikeda et al. (2013)	Winter precipitation and precipitation type	Experimental pre-HRRRv1	Eastern CONUS	30 Jan–31 Mar 2011
Pinto et al. (2015)	Warm season VIL objects	Experimental pre-HRRRv1	CONUS	JJA 2012, 2013
Bytheway and Kummerow (2015)	Warm season QPF objects	Experimental HRRRv1	Central CONUS	MJJA 2013
Griffin et al. (2017a)	Simulated satellite BTs	Experimental HRRRv2	Central and eastern CONUS	23–24 Jul 2015
Griffin et al. (2017b)	Simulated satellite BTs	Experimental HRRRv2	CONUS	August 2015 and January 2016
Ikeda et al. (2017)	Surface precipitation type	Experimental HRRRv1-v2	Eastern CONUS	DJF 2013/14, 2014/15
Bytheway et al. (2017)	Warm season QPF objects	Experimental HRRRv1-v2	Central CONUS	JJA 2013, 2014, 2015
McCorkle et al. (2018)	Surface weather conditions	Experimental HRRRv2-v3	Alaska	December 2016–June 2017
Lee et al. (2019)	Surface energy budget	HRRRv2	Northern Alabama	September 2016–April 2017
Pichugina et al. (2019)	Winds in complex terrain	HRRRv1-v2	Columbia River Gorge	2016
Radford et al. (2019)	Banded snowfall	HRRRv1-v2	Central and eastern CONUS	NDJFM 2015/16, 2016/17, 2017/18
Blaylock and Horel (2020)	Lightning threat	HRRRv2-v3	CONUS	MJJAS 2018, 2019
Fovell and Gallagher (2020)	Surface and PBL evolution	HRRRv3	CONUS	15 Dec 2018–15 Jul 2019
Duda and Turner (2021)	Convective storm objects	HRRRv3	Central and eastern CONUS	April–September 2019
Dougherty et al. (2021)	West Coast precipitation	HRRRv3	California	December 2018–February 2019
English et al. (2021)	West Coast precipitation	HRRRv3 and retro HRRRv3	California	February–March 2019
Chow et al. (2022)	Smoke forecasts	Retro HRRRv4	Western CONUS	November 2018 (Camp Fire, California)

and dewpoint, and 10-m wind speed), and cloud ceiling and visibility forecasts. The final section discusses conclusions, and outlines future directions for convection-allowing modeling in the post HRRRv4 era.

2. Verification of HRRRv1–HRRRv4

This section provides a summary of quantitative forecast performance by the HRRR system, as implemented in versions 1–4 (see D22, their Table 2). Forecast performance exhibits distinct diurnal and seasonal cycles, and performance also varies with forecast lead time. It is beyond the scope of this study to present verification for all variables across the seasonal and diurnal cycles and with forecast lead time, but we present a subset of these results that illustrate the main improvements with each subsequent version of the HRRR.

Results are presented based primarily on real-time verification statistics for the CONUS-domain experimental HRRR, run at NOAA/GSL since 2010 (see D22, their Table 2), and the operational HRRR, run at NCEP Central Operations (NCO) since 2014. Results from the experimental HRRR are

included because the period of record extends back well before the operational implementation of HRRR in September 2014, allowing us to include an earlier preoperational version of the HRRR in our analysis. While controlled retrospective experiments have been carried out and evaluated for each operational upgrade of the HRRR, these multi-season experimental retrospective periods are no more than a month each in length, and they do not overlap, which would greatly increase the number of figures required to illustrate changes between HRRR versions. While showing "unmatched" real-time verification for various versions of the HRRR is strictly not a controlled comparison, using a long time period (i.e., an entire warm season, 15 April–15 October) for each version ensures that the results are sampling the same seasonal cycle, despite some differences in mean meteorological conditions from year to year. Throughout the remainder of this article, "warm season" refers to the period 15 April–15 October, while "cool season" refers to the period 15 October–15 April. To confirm that the unmatched year-to-year comparisons between HRRR versions are not unduly affected by changes in meteorology from year to year, we examined multiple years per HRRR version where

TABLE 2. Contingency table showing the four components based on observed/forecasted events (defined as exceedance of a threshold).

	Event observed? Yes	Event observed? No
Event forecasted? Yes	a (hit)	b (false alarm)
Event forecasted? No	c (miss)	d (correct negative)

available (not shown). Sample sizes were also assessed for all of the comparisons included (see upcoming figures). The diurnal cycle of model performance is generally assessed using 6-h forecasts, because of the original design of the HRRR to provide guidance for 0–15-h forecast ranges, although longer forecasts are included where possible. Note that the sample size for longer forecasts becomes progressively smaller, due both to extended forecasts only being launched four times per day, as well as increased likelihood of abbreviated forecasts (for the experimental HRRR) and outages in the verification system ingest.

The time periods included in the experimental HRRR evaluation are intended to reflect an approximate correspondence to operational HRRR versions, and to allow some general conclusions about version-to-version changes. During the HRRR development process, controlled experiments on individual system changes were very important for evaluating candidate changes, with decisions guided by the scientific method (formulating a hypothesis and testing it, modifying the hypothesis if necessary, further testing, etc.). However, it is beyond the scope of this paper to describe all of the many changes involved in each HRRR upgrade. Instead, we take advantage of relatively "frozen" code periods, generally during the warm seasons of each year, to illustrate the net forecast performance changes associated with each upgrade. The configuration of the experimental HRRR does not always exactly match an operational HRRR version, but close correspondence occurs during the "frozen" code periods leading up to each implementation. For this reason, comparisons during the cool season are generally restricted to the operational HRRR.

Some results are also shown from controlled retrospective experiments comparing HRRR versions for a particular season. These experiments are configured identically to the real-time simulations, although not subject to the occasional interruptions and outages impacting the experimental HRRR. The retrospective runs included running the upstream RAP and HRRRDAS systems, where needed for initial and boundary conditions and ensemble covariance information.

a. Verification approach

In this article, we use the Model Analysis Tool Suite (MATS; Turner et al. 2020) developed within the NOAA Global Systems Laboratory to evaluate HRRR forecasts. Verification is undertaken in two basic frameworks: station-based (for surface weather observations), and grid-to-grid (for comparing HRRR forecasts against gridded observations of precipitation or composite radar reflectivity). For continuous variables, model errors are quantified in terms of root mean squared error (RMSE; Wilks 2011), while mean forecast biases are also calculated. For discontinuous (e.g.,

precipitation) or impact-based variables (e.g., radar reflectivity, ceiling, and visibility), verification is done in a contingency table framework in which binary events are defined (generally an exceedance of a threshold–for example, radar reflectivity greater than 35 dBZ), and gridpoint forecasts can be classified as hits, misses, false alarms, or correct negatives (Table 2). Various performance metrics can be calculated based on these contingency tables; the metrics used in this article are shown in Eqs. (1)–(4):

$$CSI = \frac{a}{a + b + c}, \quad (1)$$

$$frequency\ bias = \frac{a + b}{a + c}, \quad (2)$$

$$POD = \frac{a}{a + c}, \quad (3)$$

$$FAR = \frac{b}{a + b}. \quad (4)$$

Particular attention is required for the verification of discontinuous fields such as precipitation, in which an inherent scale mismatch exists between point surface rain gauge observations and a numerical model gridpoint (e.g., Tustison et al. 2001; Mittermaier 2014). Statistical comparisons must take into account this scale difference; similarly, comparisons between models and gridded observations must also account for differences in scale. This scale mismatch is particularly pronounced during the summer season when isolated convective rainfall is widespread, while during the cool season additional challenges emerge related to the measurement of snowfall (e.g., Randriamampianina et al. 2021).

While alternative approaches for objective verification targeting specific phenomena have been developed (e.g., object-oriented approaches for comparing model QPF with quantitative precipitation estimates (QPE), Bytheway and Kummerow 2015; Bytheway et al. 2017; surrogate severe verification, Sobash et al. 2011), these approaches are considered beyond the scope of this work.

b. Precipitation forecast accuracy (simulated composite radar reflectivity and QPF)

One of the key capabilities of the HRRR is the prediction of precipitation, and particularly precipitation associated with deep, moist convection (e.g., D22), the occurrence of which has large impacts for society. In this section, we present quantitative verification of simulated composite radar reflectivity and QPF. Simulated composite radar reflectivity, which is calculated based on model microphysical hydrometeor populations,
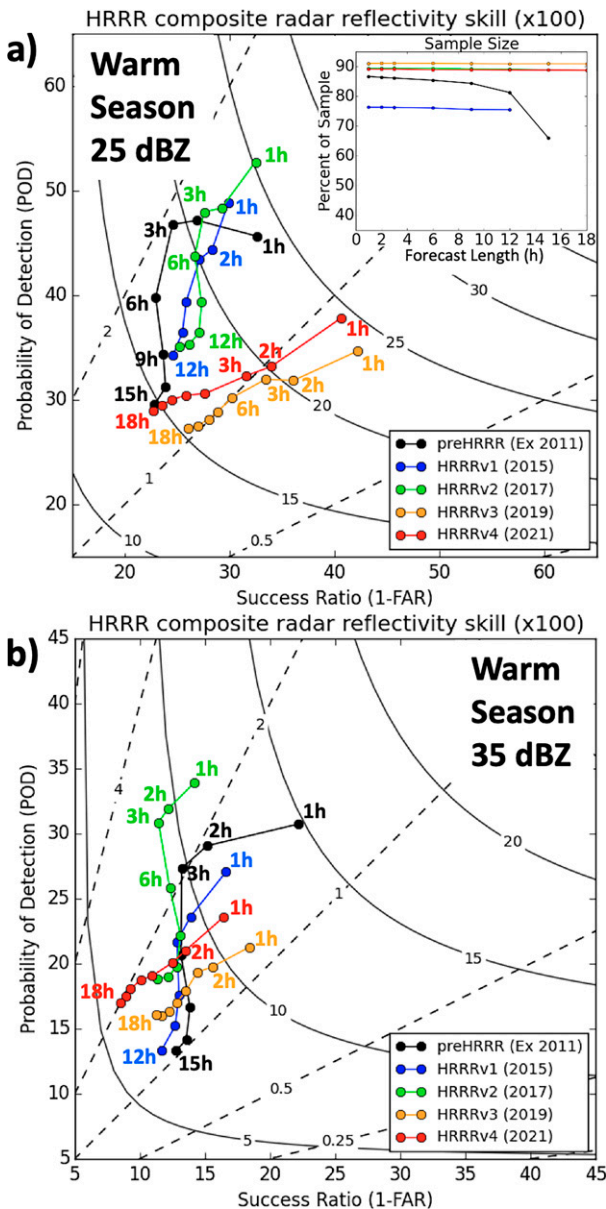
FIG. 2. Performance diagrams for HRRR composite reflectivity forecast skill by lead time for different versions of HRRR, analyzed by warm seasons (15 Apr–15 Oct). "Ex" in the legend denotes statistics from the experimental HRRR; other periods are taken from the operational HRRR. Verification is against the MRMS reflectivity mosaic over the eastern CONUS (east of 100°W longitude), using a threshold of (a) 25 and (b) 35 dBZ. The vertical axis is for probability of detection (POD), and the horizontal axis is the success ratio [defined as one minus the false alarm ratio (FAR)], both upscaled to a 20-km grid. Shown are forecast lengths at 1-h intervals (circles) to 3 h, and then every 3–18 h, where available, moving from top right to bottom left. Note that forecasts beyond 15 h are not available until HRRRv2. The inset curves in (a) shows the percent of hourly forecasts available for each time period. The eastern CONUS contains 14 554 20-km grid boxes.

is often used by forecasters to anticipate the timing, mode, and evolution of convection, and thus represents an important forecast variable. Reflectivity is verified against the Multi-Radar Multi-Sensor (MRMS) mosaic developed by NSSL (Zhang et al. 2016; Smith et al. 2016). QPF is compared against the Stage-IV QPE product; the Stage-IV QPE is interpolated to the 3-km HRRR grid using a neighborhood-budget approach. Our ability to predict fine-scale precipitation in regions of complex terrain may actually exceed our ability to observe it (Lundquist et al. 2019), so we focus our analysis on the eastern United States where there are fewer problems related to sparse rain gauges and radar beam blockage. We investigate reflectivity and QPF, as they are somewhat complementary; reflectivity is an instantaneous quantity, dependent on resolved microphysical processes, while QPF is integrated across time.

Figure 2 shows verification of 25- and 35-dBZ composite radar reflectivity forecasts over the eastern United States in performance diagrams (Roebber 2009), where all 24 daily HRRR initializations are aggregated together. Perfect forecasts appear at the upper right of the diagrams, with high probability of detection (POD) and low false alarm ratio (FAR), and with a frequency bias of one falling near the diagonal line. Furthermore, the CSI is represented by the curved lines. HRRR forecasts and MRMS analyses are upscaled to a 20-km grid for evaluation, which reduces penalizing slightly misplaced reflectivity objects in the HRRR forecasts (Mittermaier 2014). The upscaling is carried out with a neighborhood-budget approach with all the 3-km grid boxes in each 20-km area. Using an evaluation grid coarser than 20-km increases overall skill, but relative performance of the different HRRR versions is insensitive to grid size.

The pre-HRRRv1 version shown in the black curves did not carry out any direct radar DA, leading to a sharp increase in frequency bias during the early forecast hours for both reflectivity thresholds (Fig. 2). The first operational version of the HRRR (Fig. 2, blue curves) featured a 1-h "pre-forecast" hour with 15-min radar DA (described in more detail by D22), which reduced this unphysical bias maximum at 2–3-h forecast length.

The HRRRv2 (Fig. 2, green curves) featured an updated version of the Thompson microphysics scheme, as well as updated physics to combat a summertime low-level warm and dry bias which was leading to spurious convective initiation over the central United States (see section 2c for further details). These improvements resulted in higher CSI scores throughout the forecasts, with small improvements in frequency bias at longer forecast lengths for the 25-dBZ threshold (Figs. 2a,b). HRRRv3 (Fig. 2, orange curves) featured further DA and physics changes. The reduction in frequency bias from HRRRv2 to HRRRv3 is largely attributable to the reduced latent heating applied during the RAP diabatic digital filter initialization (DDFI) from RAPv3 to RAPv4 (see Weygandt et al. 2022, their Table 2).

The final version of HRRR, HRRRv4, features ensemble radar DA based on the 36-member HRRRDAS (D22, section 3d therein), in addition to significant physics improvements. This
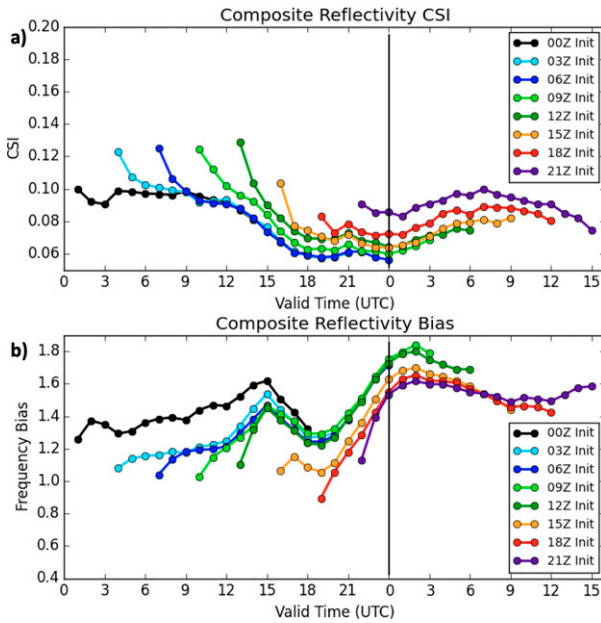
FIG. 3. Time series of (a) critical success index (CSI) and (b) frequency bias of 1–18-h forecasts of 35-dBZ composite radar reflectivity over the eastern CONUS for the experimental HRRRv4 during warm season (15 Apr–15 Oct) 2020. Shown are HRRR initializations every 3 h during the day, with the x axis showing valid time in UTC. The CSI is upscaled to a 20-km grid.

results in radar reflectivity forecasts that have a relatively unchanging bias with increasing forecast length. Overall, HRRRv4 reflectivity forecasts have a higher bias than HRRRv3, which is an improvement for 1–3-h forecasts of 25 dBZ, but a degradation at longer ranges and at 35 dBZ.

While Fig. 2 shows performance diagrams including all hourly initialization times, it is possible to stratify results by diurnal initialization time. Figure 3 shows time series of CSI and frequency bias for prediction of 35 dBZ composite radar reflectivity for the experimental HRRRv4 during the warm season of 2020. In general, the benefit of DA can be seen with the relatively high CSI (Fig. 3a) and reduced high frequency bias (Fig. 3b) during the first few hours of each subsequent forecast. However, all forecasts valid during the 1500–0300 UTC period exhibit a dip in CSI due to the challenges of predicting convective initiation (CI) and early convective evolution. In general, HRRRv4 predicts too much convection (Fig. 2b), but the shape of the bias curves (Fig. 3b) demonstrates that the model has challenges with the timing of CI. The bias reaches a minimum near 1900 UTC for all initializations, followed by a maximum near 0200 UTC, indicating that the model overpredicts the upscale growth of convection after CI.

Figure 4 illustrates QPF performance through the various HRRR incarnations, as compared against Stage-IV QPE. This figure compares 6-h experimental HRRR QPF against 6-h Stage-IV QPE for (Fig. 4a) warm seasons and (Fig. 4b) cool seasons. Note that year-to-year variability becomes significant for the relatively rare events of greater than 1 in. (6 h)$^{-1}$, so
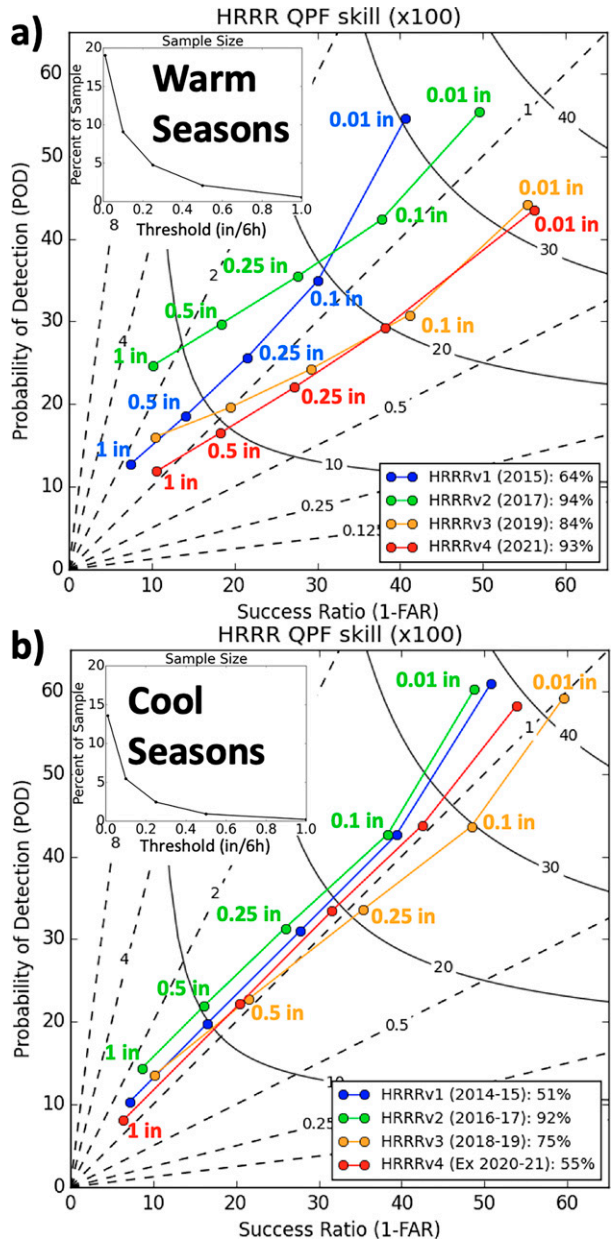

FIG. 4. Performance diagrams for HRRR QPF skill by precipitation threshold for different versions of the HRRR, analyzed by (a) warm seasons (15 Apr–15 Oct) and (b) cool seasons (15 Oct–15 Apr). "Ex" in the legend denotes statistics from the experimental HRRR; other periods are taken from the operational HRRR. Verification is against the 6-h Stage-IV QPE over the eastern CONUS (east of 100°W longitude), evaluating 0–6-h experimental HRRR QPFs against 6-h Stage-IV QPE. The vertical axis is for POD, and the horizontal axis is the success ratio (1 − FAR), both upscaled to a 20-km grid. Shown are 0.01-, 0.1-, 0.25-, 0.5-, and 1-in. thresholds, moving from top right to bottom left. Percentage values in the legend indicate the data availability for each time period. The inset curves show the percent of 6-h QPEs exceeding each threshold for (a) 15 Apr–15 Oct 2017 and (b) 15 Oct 2016–15 Apr 2017. The eastern CONUS contains 14 034 20-km grid boxes.
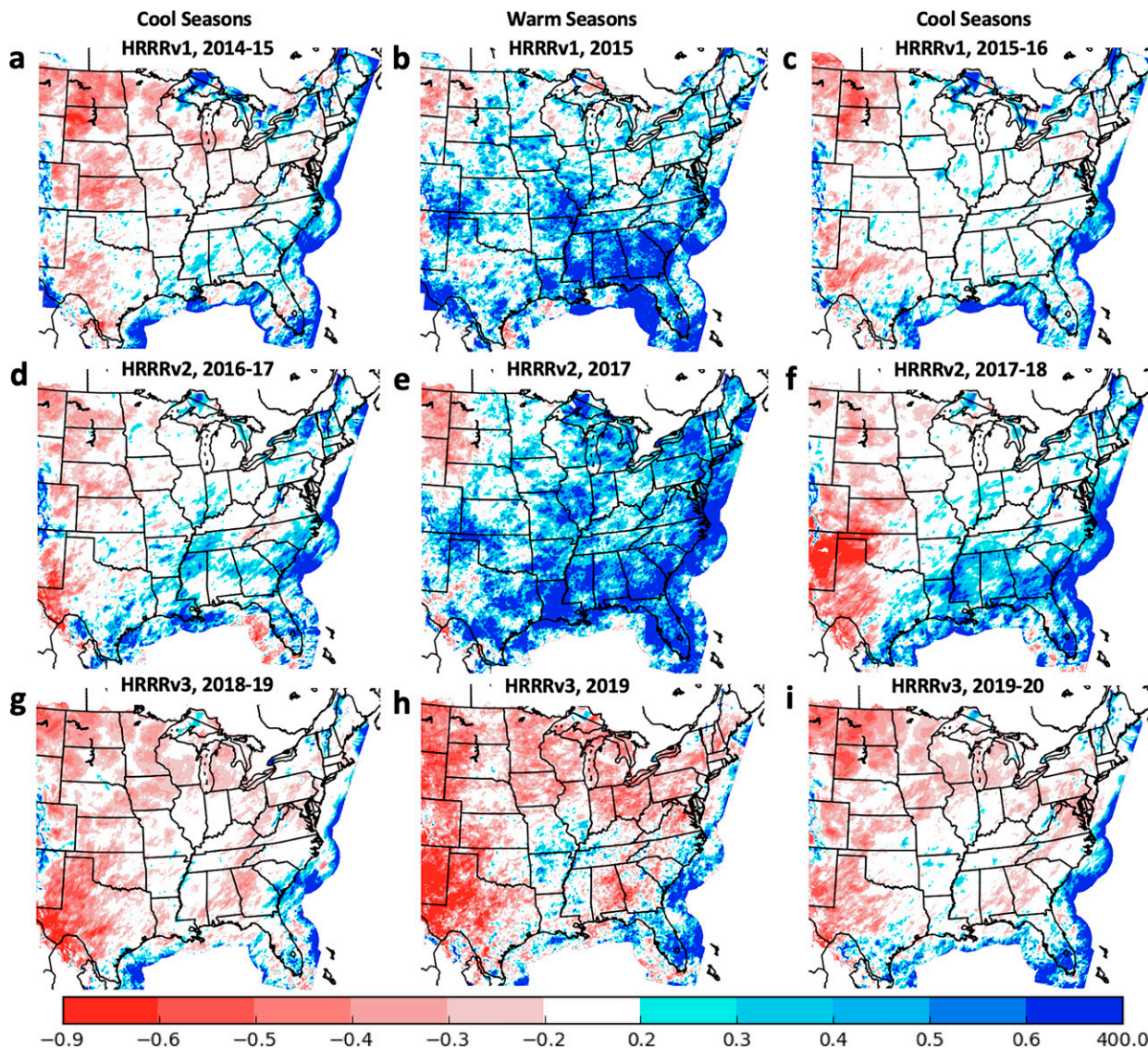
FIG. 5. Maps of the ratio of total 0–6-h HRRR QPFs vs Stage-IV 6-h QPEs for operational (a)–(c) HRRRv1, (d)–(f) HRRRv2, and (g)–(i) HRRRv3. The ratio is formulated as (QPF–QPE)/QPE, such that zero indicates perfect agreement for the season; positive values indicate QPF > QPE, and negative values indicate QPF < QPE. Shown are (left),(right) cool seasons (15 Oct–15 Apr) and (center) warm seasons (15 Apr–15 Oct).

we exclude thresholds heavier than this. The HRRRv1 exhibits a monotonic decrease in QPF skill with increasing precipitation threshold, with a high bias evident at all thresholds during the warm season (Fig. 4a, blue curve); HRRRv2, benefiting from model physics changes, exhibits a dramatically higher CSI than HRRRv1 [note the position of the 0.25, 0.5, and 1 in. $(6 \text{ h})^{-1}$ points in Fig. 4a, blue curve to green curve] but exhibits a greatly increased frequency bias relative to HRRRv1 (Fig. 4a, green curve). The HRRRv3 and HRRRv4 have exhibited a low bias for light to moderate precipitation amounts (less than one inch per 6 h), although less so at longer forecast lead times (not shown). He et al. (2022, manuscript submitted to *J. Adv. Model. Earth Syst.*) also document a HRRRv4 low bias in 2-m moisture and latent heat flux due to

a dry bias in soil moisture related to an underprediction of precipitation. Candidates for reduced 0–1-h precipitation in HRRRv4 are related to an aspect of its data assimilation (use of ensemble mean) and cooler 2-m temperatures. Overall, HRRRv2 exhibits the highest skill for warm season QPF, but at the cost of a high bias at moderate to high precipitation thresholds.

For the cool season (Fig. 4b), differences between HRRR versions appear much smaller than in the warm season, although skill is generally higher than during the warm season. Changes in HRRRv3 and HRRRv4 led to reductions in the high wintertime QPF frequency bias at the lower thresholds.

Figure 5 shows seasonal evaluations of 0–6-h HRRR QPF versus Stage-IV QPE for the various versions of the HRRR. The maps are presented in terms of the difference (QPF minus QPE)
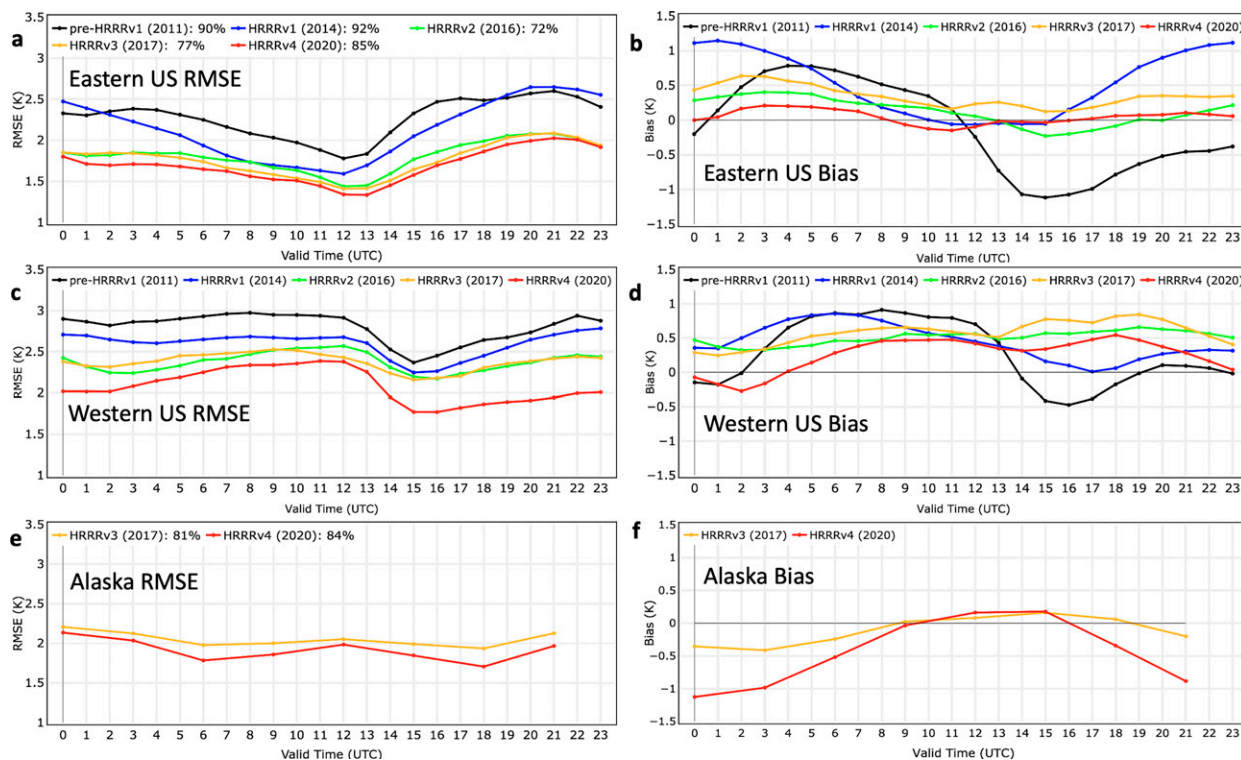
FIG. 6. The 2-m temperature 6-h forecast performance by valid time for the experimental HRRR, analyzed by warm seasons (15 Apr–15 Oct), which approximately correspond to operational versions of the HRRR. Verification is against METAR observations over (a),(b) the eastern CONUS (generally 1380–1480 sites); (c),(d) the western CONUS (generally 510–580 sites); and (e),(f) Alaska (generally 100–130 sites). (left) RMSE and (right) bias. Percent of hourly 6-h forecasts available for each period are indicated in the legend in (a) for CONUS and in (e) for Alaska.

normalized by QPE, where both QPF and QPE are summed over the periods indicated. Due to the timing of HRRR implementations during the summer, there are more unbroken cool seasons than warm seasons of a fixed HRRR version available for the evaluation. The left and right columns show cool seasons for various years, while the middle column shows warm seasons. HRRRv1 is shown in the top row, HRRRv2 in the middle row, and HRRRv3 in the bottom row; the western United States is excluded from the analysis due to the uncertainties with Stage-IV QPE in that region. HRRRv4 results are not shown in this figure as it has not yet run long enough to provide seasonal results.

The availability of two cool seasons for each HRRR version allows a qualitative assessment of the statistical significance of the regional features which emerge. It is seen that each HRRR version exhibits consistent patterns of disagreement with QPE across two different cool seasons. HRRRv1 exhibits a relative dry bias in the northern plains, but HRRRv1 QPF agrees relatively well with QPE in the southeastern CONUS (Figs. 5a,c). HRRRv2 exhibits a similar pattern, but with a moist bias in the southeastern CONUS (Figs. 5d,f). HRRRv3 once again has a more neutral bias in the southeastern CONUS, but with a more pronounced dry bias in the high plains and the northern CONUS. Evaluating 6–12-h QPFs (not shown) reveals similar spatial patterns, although drier (closer to QPE) in the southeastern CONUS, particularly for HRRRv1 and HRRRv2.

During the warm season (Figs. 5b,e,h), HRRRv1 and HRRRv2 both exhibited a significant moist bias throughout the eastern CONUS (Figs. 5b,e). HRRRv3 exhibits a dramatic improvement in the warm-season QPF, with a more neutral bias in the southern and eastern CONUS (Fig. 5h), although introducing a dry bias in the northern CONUS. These results are subject to uncertainties related to the quality of the Stage-IV QPE dataset, but they allow some generalizations regarding changes in HRRR 0–6-h QPF between versions. At longer forecast ranges (6–12 h; not shown), the HRRRv1 and HRRRv2 exhibited a much reduced moist bias throughout the CONUS. These results suggest that the initialization of convection in HRRRv1 and HRRRv2 led to excessive QPF in the first few hours of the simulation, with less impact in HRRRv3; this is again largely attributable to the reduced strength of latent heating in the RAP DDFI from RAPv3 to RAPv4 (see Weygandt et al. 2022, their Table 2).

c. Surface weather forecast accuracy (2-m temperature and dewpoint, and 10-m wind)

Since the vast majority of human activity occurs at or near Earth's surface, another key metric of weather forecast accuracy is the performance of surface weather forecasts. Temperature and moisture observations at shelter level [2 m above ground level (AGL)] can be used to evaluate HRRR forecasts, in
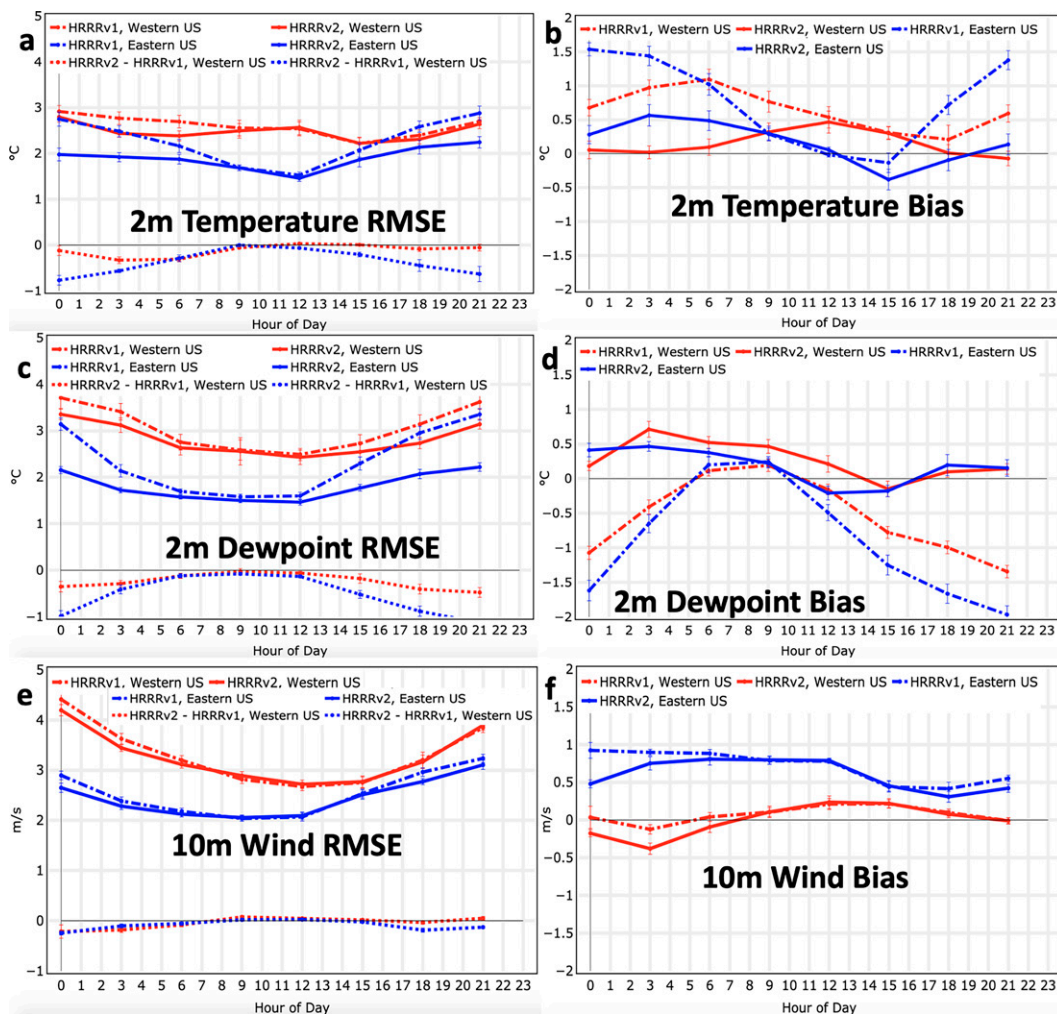
FIG. 7. The 2-m temperature 6-h forecast performance by valid time for HRRRv1 (dashed–dotted curves) and HRRRv2 (solid curves), for a summer retrospective experiment during 15 Jul–15 Aug 2014. (a),(b) The 2-m temperature; (c),(d) 2-m dewpoint temperature; and (e),(f) 10-m wind speed. (left) RMSE; and (right) bias. Results are shown for the eastern CONUS (blue) and the western CONUS (red). Difference in RMSE (HRRRv2 − HRRRv1) is shown in (a) and (c).

addition to wind observations at standard height (10 m AGL). The verification shown here is based on forecast comparisons against METAR stations in the HRRR CONUS domain (~1800 stations) mostly in the lower 48 United States and in Alaska. We present verification of 6-h forecasts, averaged at each hour of the diurnal cycle, as well as forecast performance by lead time. More detailed analysis of HRRRv3 is presented by Fovell and Gallagher (2020), and of HRRRv3 and HRRRv4 by He et al. (2022, manuscript submitted to *J. Adv. Model. Earth Syst.*). Again, Tables 2 and 4 in D22 provide an excellent reference on the changes involved in each HRRR upgrade for interpreting these results.

1) 2-M TEMPERATURE FORECASTS

Figure 6 shows the average diurnal cycle of 2-m temperature 6-h forecast RMSE (left) and bias (right) for the eastern United States (top), western United States (middle), and Alaska (bottom), for warm season 6-h forecasts. Verification for the CONUS subdomains is based on hourly HRRR forecasts, while the Alaska verification is based on HRRR-AK simulations launched only every 3 h (see D22, their Table 4). For the eastern and western CONUS, and for all HRRR versions, 2-m temperature errors are generally lowest around or shortly after sunrise (Figs. 6a,c). In terms of 2-m temperature bias, HRRRv1 and especially the pre-HRRRv1 exhibited a strong diurnal cycle of bias in both the eastern and western CONUS (Figs. 6b,d, black curves). HRRRv1 had a warm and dry bias in the daytime PBL over much of the domain during the summer, particularly in the eastern CONUS (Fig. 6b, blue curve); it was hypothesized that this bias and the associated spurious convective development mentioned in the previous section were linked by a mechanism described in the feedback cycle outlined in Fig. 10 of B16. Insufficient cloud cover in the HRRRv1 was leading to overly deep mixing and too-deep
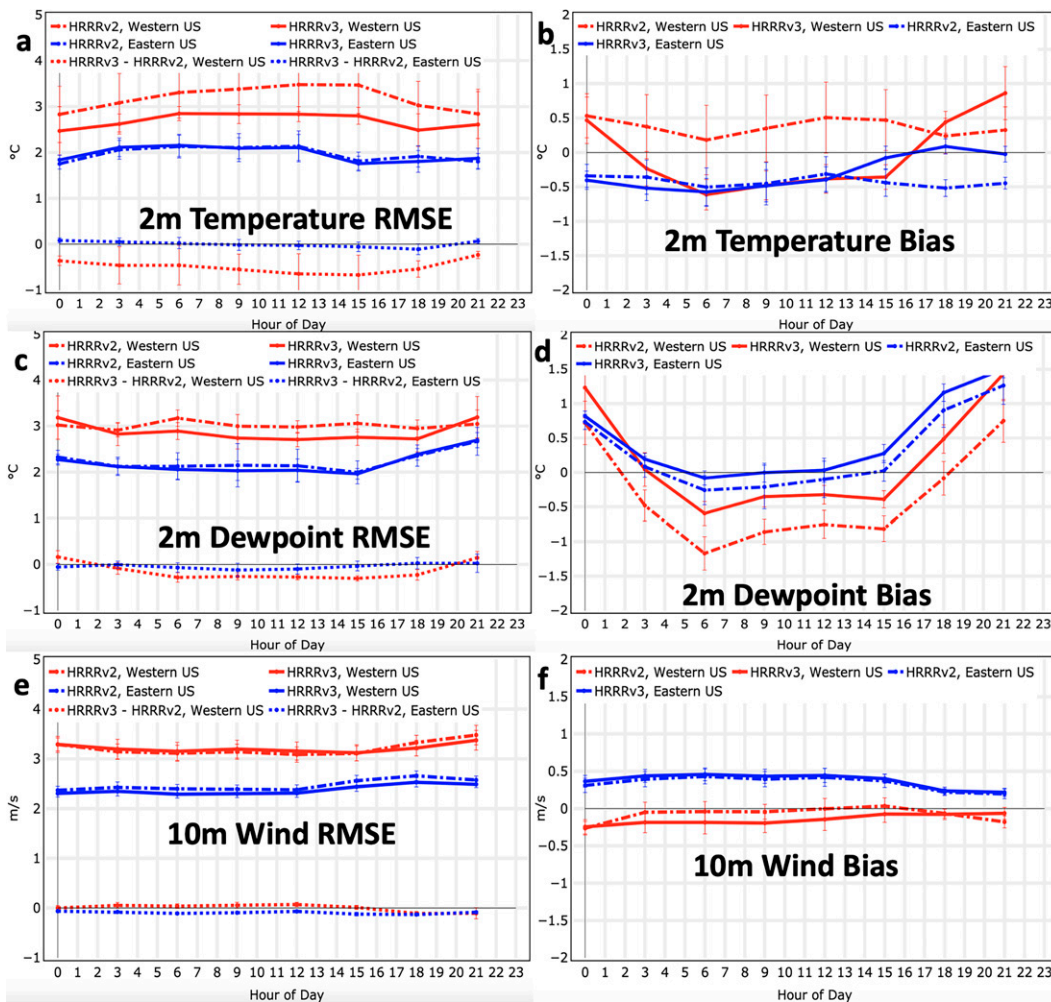
FIG. 8. As in Fig. 7, but comparing HRRRv2 and HRRRv3 for a winter retrospective experiment (1–31 Jan 2017).

PBLs, especially in the summertime, and an excess of incoming solar irradiance. This excessive low-level mixing tended to overcome convective inhibition too readily, producing spurious convection in the model.

To alleviate the biases in the initial version of the operational HRRR, development took place particularly in the model physics parameterizations. One of the foremost changes implemented in HRRRv2 was an increase of the "wilting
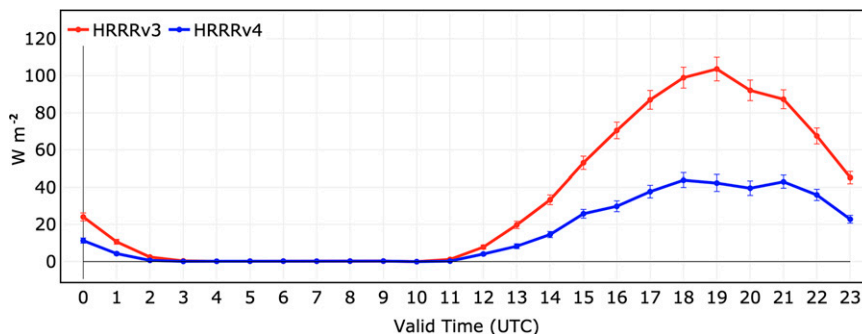
FIG. 9. Average downwelling shortwave irradiance bias for 6-h forecasts from HRRRv3 (red) and HRRRv4 (blue) from 1 Dec 2019–1 Dec 2020. Verification is against 14 SURFRAD/SOLRAD observations across CONUS.
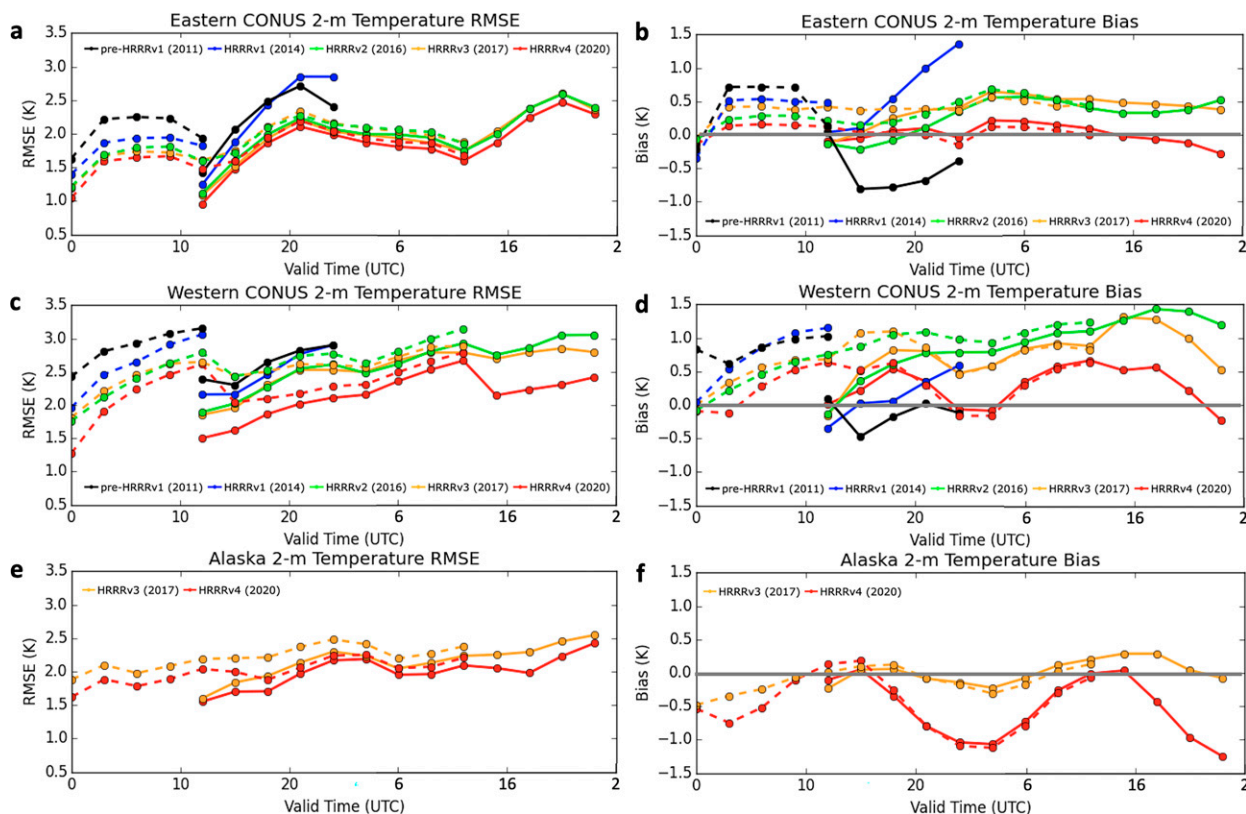
FIG. 10. As in Fig. 6 (during warm season), but by forecast lead time, and showing forecasts initialized at 1200 UTC (solid lines) and 0000 UTC (dashed lines).

point" within cropland regions in the RUC LSM, effectively allowing continued transpiration from irrigated crops and increasing low-level relative humidity; the effects of this change were most pronounced over the agriculture-rich Great Plains of the United States. Another major adjustment was allowing the RRTMG radiation scheme to interact with boundary layer clouds within the MYNN PBL scheme, having the net effect of increasing low-level cloudiness and reducing solar irradiance reaching the surface. A secondary low-level cooling effect comes from attenuation from climatological aerosol loading within the Thompson aerosol-aware microphysics scheme. These changes are described in more detail by B16 (their section 6). Within the HRRR DA, a number of changes were made to address the warm/dry bias in the HRRRv1. Hybrid ensemble-variational DA, having been shown to greatly improve forecasts of upper-level wind and other variables within the 13-km RAP (Hu et al. 2017), was implemented in the HRRR beginning with HRRRv2. Focusing more specifically on the model biases, PBL "pseudo-innovations" (B16, their section 2f) were introduced for surface temperature (in addition to surface dewpoint) in order to extend the influence of surface observations in the vertical in well-mixed situations. In addition, assimilation of 2-m temperature and dewpoint observations was modified to be more consistent (accounting for the difference in height between the typical 2-m height of sensors and the lowest model level, near 8 m AGL). These changes led to the error

characteristics outlined by the green curves (HRRRv2) in Fig. 6. HRRRv2 featured a dramatically improved 2-m temperature RMSE and bias around the clock in the eastern United States (Figs. 6a,b), and a reduction in RMSE in the western United States (Fig. 6c). Figure 7 shows results for a controlled retrospective experiment comparing HRRRv1 versus HRRRv2 performance for the summer season (15 July–15 August 2014). As shown in Figs. 7b and 7d, the daytime warm and dry bias in HRRRv1 was particularly pronounced in the summer, and was dramatically improved with the physics and DA changes implemented in HRRRv2. These improvements also led to major RMSE reductions during the daytime for both 2-m temperature and 2-m dewpoint temperature (Figs. 7a,c).

HRRRv2 exhibited a different set of biases. In particular, forecasters noted a continued tendency for the model to quickly erode low-level cloud cover. A high-frequency bias in simulated radar reflectivity and precipitation in the first few hours of the HRRR forecasts was also noted. Due to these issues, the focus of the HRRRv3 upgrade was to improve retention of low clouds, and reducing a short-lead-time high precipitation/simulated radar reflectivity bias, as well as improved 2-m temperature/dewpoint diurnal cycles in summertime. DA changes in the HRRRv3 were motivated by observed short-range forecast biases present in the HRRRv2. In particular, the high bias in precipitation during the first few hours of the forecast motivated a reduction in the strength of the latent heating applied in the
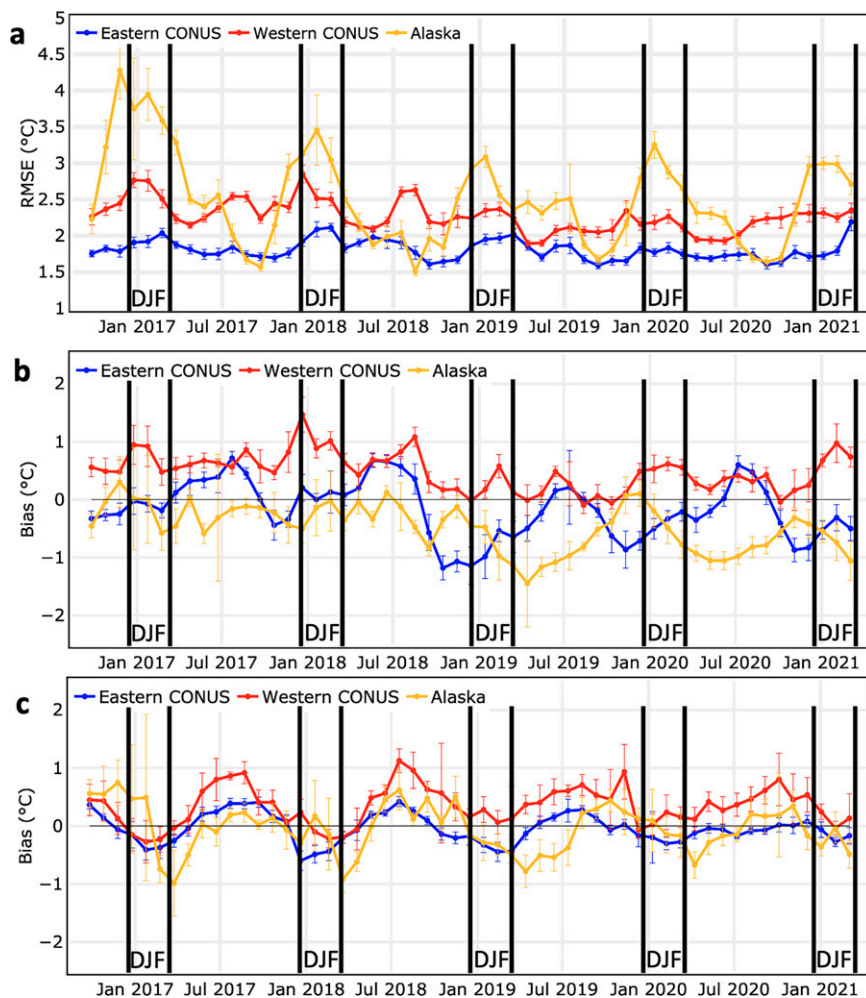
FIG. 11. The 2-m temperature 6-h forecast performance by valid time for the experimental HRRR, analyzed over a ~4.5-yr period, using 30-day averages. Verification is against METAR observations; (a) RMSE, (b) daytime bias, and (c) nighttime bias. Daytime is considered 1500–2100 UTC for the eastern CONUS, 1700–2300 UTC for the western CONUS, and 1900–0100 UTC for Alaska. Nighttime is considered 0300–0900 UTC for the eastern CONUS, 0500–1100 UTC for the western CONUS, and 0700–1300 UTC for Alaska.

RAP DDFI within regions of high observed three-dimensional radar reflectivity (see D22, their Table 2). In the realm of model physics, significant updates were made to several parameterization schemes (for more details, see D22, their Table 4 and sections 2b and 2c). HRRRv3 improvements for the surface variables are relatively small when averaged through the warm season (Fig. 6, yellow curves). However, Fig. 8 illustrates that improvements were more substantial during the winter season; these results are derived from a controlled retrospective experiment during 1–31 January 2017. Improved treatment of subgrid snow coverage, as well as subgrid cloud cover, led to significant reductions in forecast errors for 2-m temperature and dewpoint in the western United States (Figs. 8a,c), and reduction of a daytime cool bias in the eastern United States (Fig. 8b). A nighttime warm and dry bias is also reduced in the western United States (Figs. 8b,d). HRRRv3 represents the

first operational version of the HRRR, which included an Alaska domain, and Figs. 6e and 6f show error and bias characteristics of these initial Alaska forecasts (yellow curves).

The HRRRv4, as described in the previous section, featured several major advances, both for DA and model physics, leading to improved forecasts both for the CONUS and Alaska domains. A major focus of the HRRRv4 upgrade was further targeting the representation of low clouds and their tendency to prematurely erode in the HRRRv3 configuration. In addition, a storm-scale ensemble DA capability is introduced for the CONUS HRRR for the first time, as described by D22 (section 3d), resulting in improved short-range forecasts of convective evolution and the PBL. These changes led to forecast performance characterized by the red curves in Fig. 6. The HRRRv4 exhibits decreases in 2-m temperature RMSE for all three regions (Figs. 6a,c,e), but most pronounced in the western
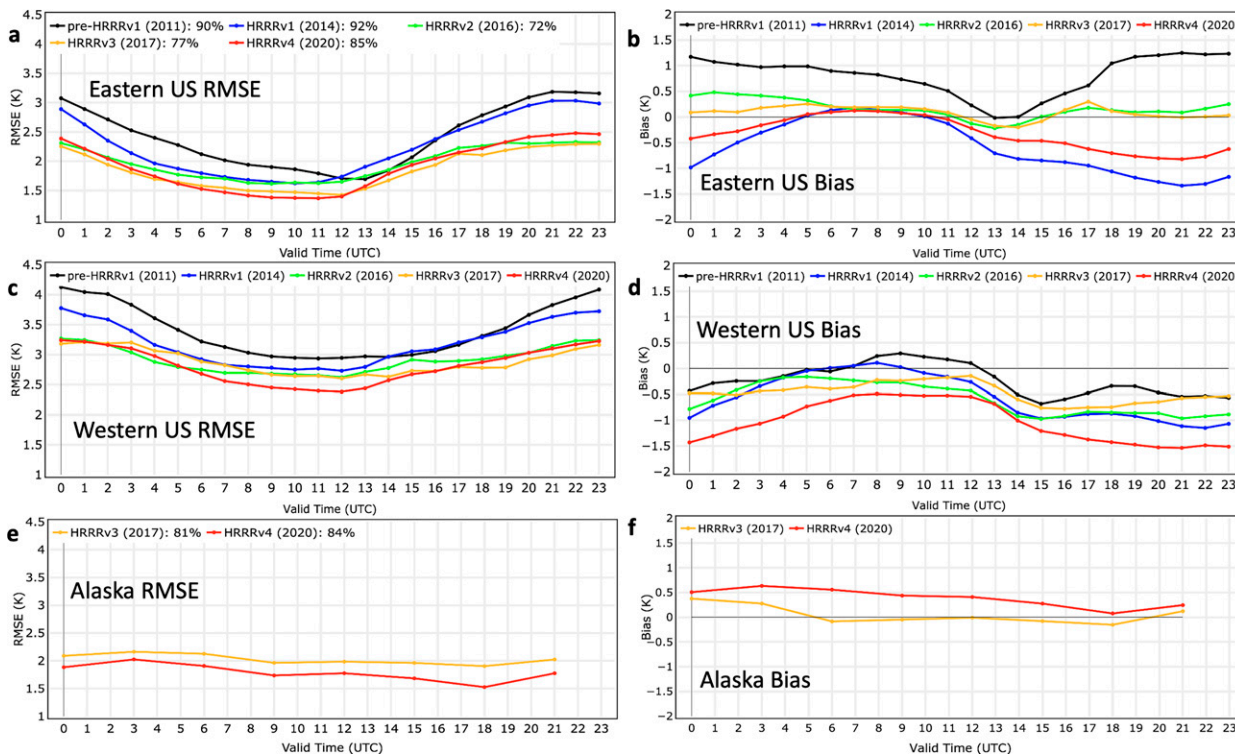
FIG. 12. As in Fig. 6 (for 6-h forecasts during warm season), but for 2-m dewpoint temperature.

United States (Fig. 6c). The dramatic improvements in cloud coverage associated with the HRRRv4 upgrade are further illustrated in Fig. 9, showing average biases in 6-h forecast shortwave irradiance across the seasonal cycle for the operational HRRRv3 versus the experimental HRRRv4 during 2020. Shortwave irradiance forecasts are evaluated against the Surface Radiation Budget (SURFRAD) and SOLRAD Network (Augustine et al. 2000), with 14 of 16 stations reporting during this period, spread across the lower 48 United States. It is evident in Fig. 9 that the high bias in downwelling shortwave irradiance is reduced by up to 50% in HRRRv4.

Figure 10 shows 2-m temperature forecast performance by lead time for the various HRRR versions, for warm season simulations initialized at 0000 (dashed lines) and 1200 UTC (solid lines). Note that Fig. 10 differs from Fig. 6 in that it shows forecast skill across lead times for just two daily model initialization times, while Fig. 6 shows 6-h forecasts from all initialization across the diurnal cycle. Across CONUS, both the preoperational version of HRRR (black curves) and HRRRv1 (blue curves) featured large RMSEs up to 12-h forecast length (Figs. 10a,c). The preoperational HRRR exhibited a nighttime warm bias and daytime cool bias (Figs. 10b,d; black curves); HRRRv1, on the other hand, exhibited a warm bias, increasing with forecast lead time (blue lines). Later HRRR versions show incremental reduction of this daytime warm bias in the eastern CONUS (Fig. 10b). The HRRRv2 (green curves) exhibited lower CONUS RMSEs out to 12-h forecast length (Figs. 10a,c), associated with the physics improvements described above. Changes in warm season 2-m temperature

RMSE and bias from HRRRv2 to HRRRv3 (yellow curves) were subtle, but improvements are seen again with HRRRv4 (red curves). In particular, the largest reductions in RMSE are seen in the western United States (Fig. 10c), associated with improved covariance representation in complex terrain, and improved treatment of subgrid clouds. Warm season results for Alaska are shown in Figs. 10e and 10f; once again we see an improvement in 2-m temperature RMSE in HRRRv4 (Fig. 10e), but an afternoon/early evening cool bias of ~1°C present in HRRRv4 (Fig. 10f).

Figure 11 illustrates the seasonal cycle of skill for 6-h forecasts of 2-m temperature from the experimental HRRR, showing RMSE (Fig. 11a), daytime bias (Fig. 11b), and nighttime bias (Fig. 11c). The Northern Hemisphere winter season (December–February) is indicated for clarity. The time series includes several HRRR versions, but consistent seasonal patterns emerge. In general, highest errors and coolest bias are seen in the winter for the eastern United States, with warm biases in the summer. In the western United States, similar patterns exist, but with larger errors than in the eastern United States, and with warmer 2-m temperature biases (warm biases overall throughout the seasonal cycle). As can be seen in Fig. 11a, Alaska is a challenging domain for short-range forecasts, partially due to the lack of observations (both locally, and upstream for the global model which provides lateral boundary conditions); errors for 2-m temperature are close to those over the eastern United States during the summer season, but much larger than either CONUS region during the winter (Fig. 11a). Large wintertime temperature errors in
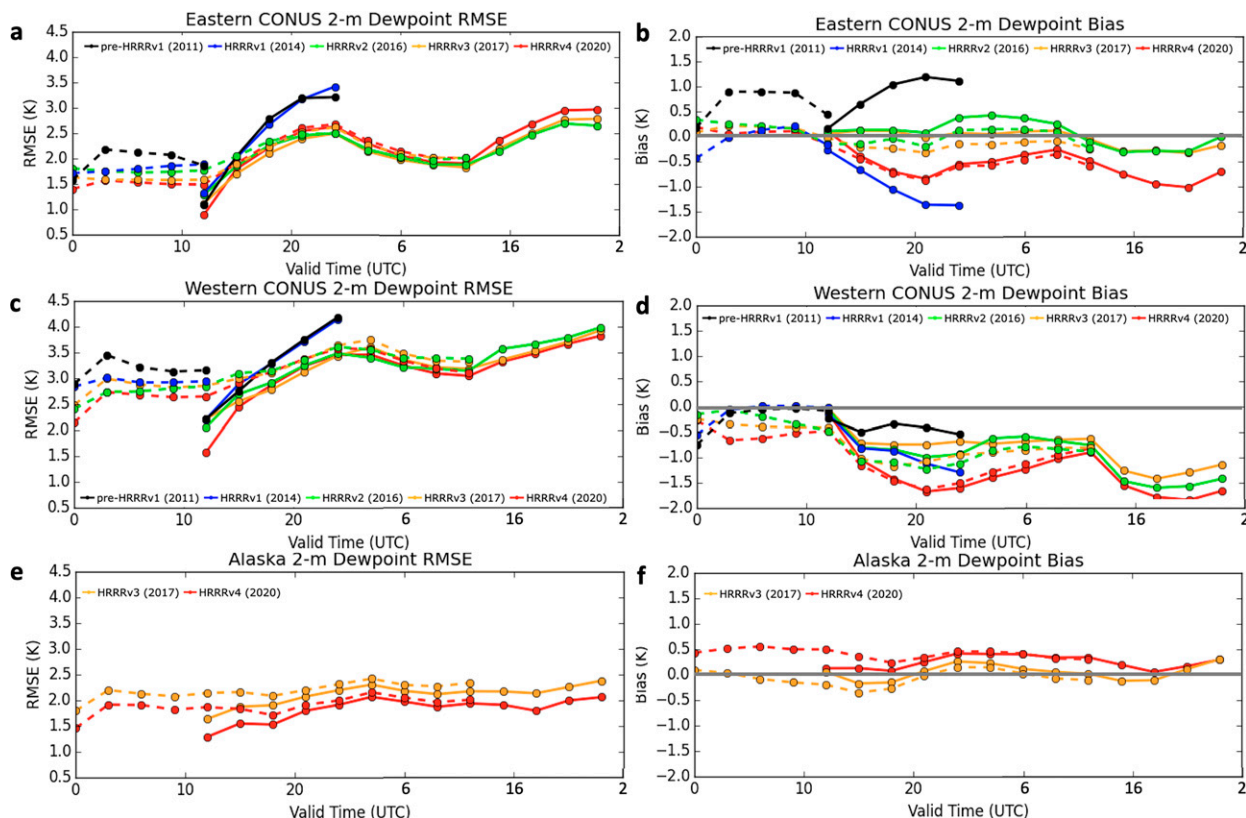
FIG. 13. As in Fig. 10, but for 2-m dewpoint temperature.

Alaska are also likely related to the development of extreme surface-based inversions and shallow arctic air. Temperature biases in Alaska are near neutral in the autumn, but quickly develop into a significant cool bias by late winter/early spring during both the daytime and the nighttime (figs. 11b,c). Note that the diurnal cycle is very muted during the Alaskan winter due to the high latitude.

2) 2-M DEWPOINT FORECASTS

Figure 12 shows 6-h forecast skill for warm season 2-m dewpoint temperature forecasts for the various model versions. Dewpoint temperature forecast errors are lowest overnight (Figs. 12a,c,e), as mixing out of low-level moisture is less prevalent at night. The pre-HRRRv1 version featured a substantial moist bias in the eastern United States, most pronounced during the afternoon and overnight hours (Fig. 12b, black curve). HRRRv1 exhibited a dramatic reduction of this moist bias (to a pronounced dry bias during the daytime), with a major decrease of dewpoint RMSE during the late afternoon and overnight hours (Figs. 12a,b, blue curves). HRRRv2, with changes aimed at improving the representation of the summertime PBL and reducing the occurrence of spurious convection, exhibited a more neutral dewpoint bias in the eastern United States during the daytime and a reduced daytime RMSE in both the eastern and western United States (Figs. 12a,b,c, green curves). As for 2-m temperature forecasts, warm season

HRRRv3 performance (shown by the orange curves in Fig. 12) is similar to HRRRv2. HRRRv4 exhibits relatively low 2-m dewpoint RMSEs for all three domains (Figs. 12a,c,e; red curves), but a dry bias for the CONUS, most pronounced in the western United States (Figs. 12b,d). The summertime dewpoint bias over Alaska is near neutral (Fig. 12f).

Figure 13 shows warm season forecast performance by lead time for 2-m dewpoint temperature. Dewpoint RMSEs exhibit a stronger increase with lead time in 1200 UTC initializations as compared with 0000 UTC initializations, particularly over the CONUS (Figs. 13a,c). Dewpoint RMSEs exhibit major nighttime reductions from pre-HRRRv1 to HRRRv1, and daytime reductions from HRRRv1 to HRRRv2 (Figs. 13a,c). In the eastern CONUS, the preoperational HRRR exhibited a moist bias throughout the diurnal cycle, but HRRRv1 exhibited a daytime dry bias (increasing with lead time) but a near neutral bias in nighttime (Fig. 13b). HRRRv2 and v3 exhibit near neutral 2-m dewpoint biases in the eastern United States (Fig. 13b), with minimal bias growth during the forecast, while HRRRv4 exhibits an increasing daytime dry bias in the CONUS (Figs. 13b,d). Most versions of HRRR exhibit a dry bias in the western CONUS (Fig. 13d), which, in combination with the daytime warm bias (Fig. 10d), has been noted by forecasters to contribute to predictions of worse fire weather conditions as compared with other forecast guidance. For Alaska, HRRRv4 has reduced 2-m dewpoint RMSEs, and a near neutral bias (Figs. 13e,f, red curves).
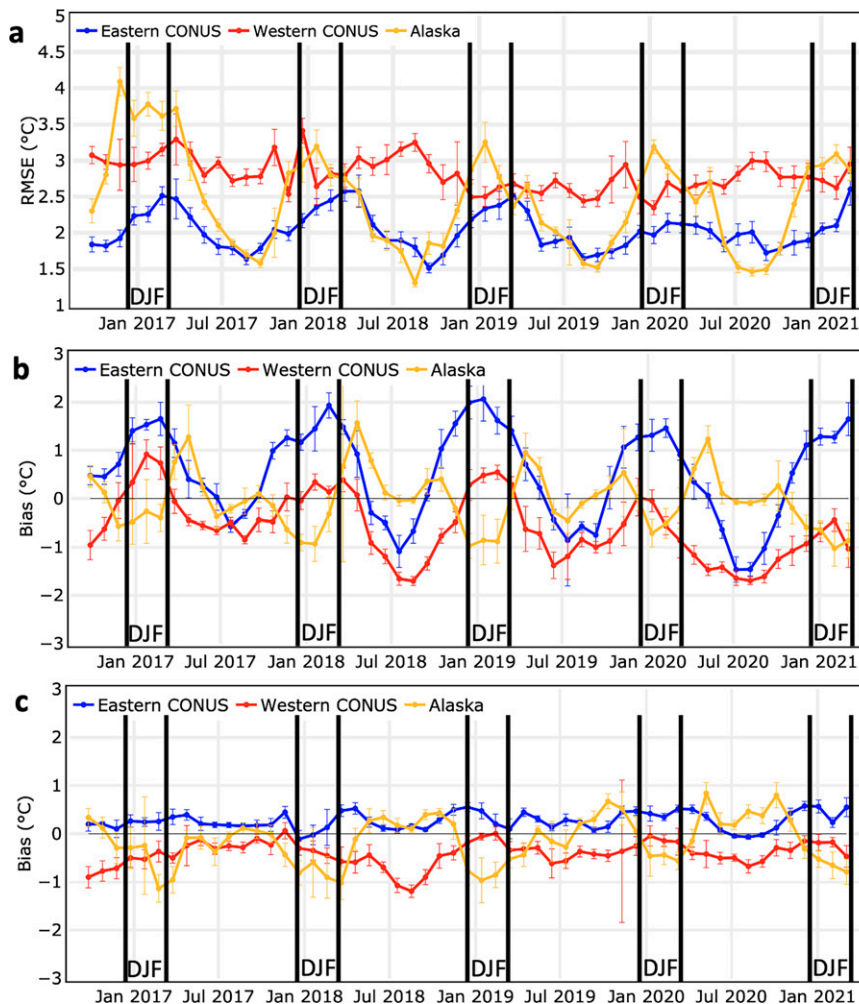
FIG. 14. As in Fig. 11, but for 2-m dewpoint temperature.

Figure 14 shows the seasonal cycle of 2-m dewpoint temperature errors over the past few years. Overall, seasonal cycles of errors for the various domains are similar to those for 2-m temperature (Fig. 11), although errors for dewpoint tend to be higher, particularly in the western United States. Daytime dewpoint biases tend to be somewhat anticorrelated with 2-m temperature biases across the seasonal cycle; e.g., warm biases are associated with dry biases, and vice versa. Daytime dewpoint biases in the CONUS exhibit a strong seasonal cycle with a wintertime moist bias and a summertime dry bias. In the eastern United States, 2-m dewpoint forecasts exhibit a larger-magnitude seasonal cycle in RMSE than 2-m temperature forecasts.

3) 10-M WIND FORECASTS

Figure 15 shows warm season 6-h forecast performance for 10-m wind speed. Forecast skill for 10-m winds has improved from HRRR version to HRRR version, with reductions in RMSE for all three domains (Figs. 15a,c,e). Wind speed bias exhibits much less version-to-version variability than 2-m

temperature or dewpoint. Wind speed forecasts exhibit a somewhat high bias (Figs. 15b,d,e), particularly in the eastern United States (Fig. 15b). HRRRv4, featuring a gravity wave drag scheme (e.g., D22, section 2b therein) has lower RMSE in the western United States than HRRRv3 (Fig. 15c); the improvement is more minor in the eastern United States due to the smoother terrain in that region (Fig. 15a). The HRRRv4 for Alaska does not run a gravity wave drag scheme (see D22, p. 15).

Figure 16 shows warm season forecast performance by lead time for 10-m wind speed. Similarly to 2-m dewpoint forecasts (Fig. 13), errors increase much more rapidly in 1200 UTC initializations than in 0000 UTC initializations. Successive versions of HRRR generally feature reduced 10-m wind speed RMSE (Figs. 16a,c,e). In the eastern United States, all versions have featured a high wind speed bias (Fig. 16b). HRRRv4 has higher wind speeds than HRRRv3 for both CONUS and Alaska (Figs. 16b,d,e).

Figure 17 shows the seasonal cycle of forecast performance for 10-m wind speed. Both the eastern and western U.S.
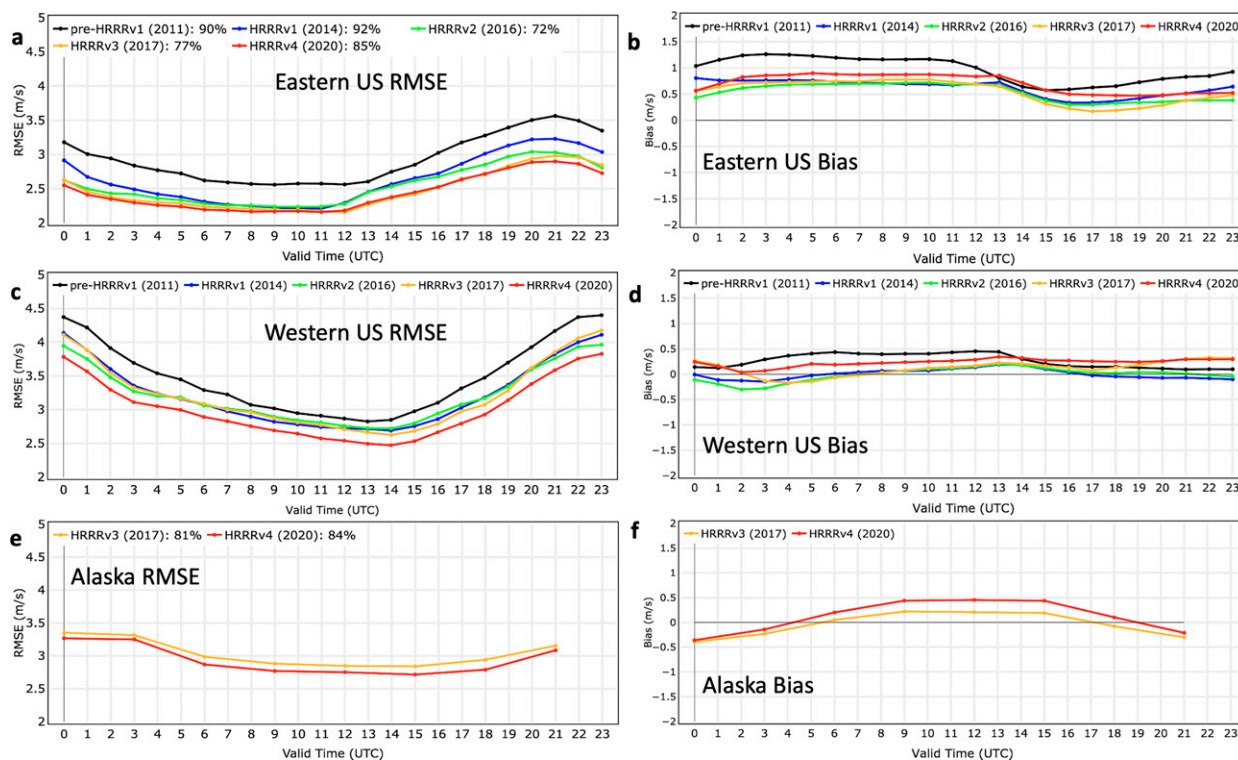
FIG. 15. As in Fig. 6 (for 6-h forecasts during warm season), but for 10-m wind speed.

domains exhibit relatively high errors during the windy spring season, but lower errors in the autumn (Fig. 17a); errors are lower in the eastern United States than the western United States. Alaska exhibits a much stronger seasonal cycle in forecast skill, with large errors in winter. HRRR forecasts for the eastern United States exhibit a consistent high wind speed bias during both day and night, while forecasts for the western United States are relatively unbiased on average (Figs. 17b,c). Wind speed forecast biases in Alaska have a seasonal cycle, with a low bias in the late winter–early spring, and a high bias in the late summer–early autumn (Figs. 17b,c), and a higher bias at night than during the daytime.

In summary, each subsequent version of the HRRR has featured targeted DA and model physics improvements aimed at least partially at improving surface weather forecasts, verified here by comparison against METAR 2-m temperature and dewpoint and 10-m wind speed observations for both the eastern and western United States. Challenges remain with the treatment of the evening transition of the PBL and other aspects of PBL evolution, but the performance of the operational HRRRv4 represents an initial baseline for evaluating next-generation convection-allowing models.

d. Cloud ceiling and surface visibility forecast accuracy

Accurate forecasts of low cloud ceilings and reduced visibility are critical for many applications, particularly for transportation. Low cloud ceilings determine flight rules for aviation and can significantly affect operations at airports. An accurate representation of cloud characteristics, including cloud ceiling,

is important for representing boundary layer evolution overall, with associated impacts on forecasts for severe weather and renewable energy, for example. HRRR carries out a unique nontraditional DA technique designed to merge cloud information from a model background with updated information from surface ceilometer observations and satellite cloud top observations. This technique, described in more detail by Benjamin et al. (2021), initializes stratiform clouds in HRRR, permitting more accurate forecasts in the first few hours after initialization. In this section, we quantify forecast performance in terms of prediction of the occurrence of cloud ceilings below the key thresholds of 3000, 1000, and 500 ft AGL (914, 305, and 152 m AGL), and the occurrence of surface visibility below the thresholds of 5, 3, and 1 mi (8, 4.8, and 1.6 km).

Figure 18 describes HRRR cloud ceiling forecast performance across various versions. These results are taken from the operational HRRR and are for years defined as December–November, allowing comparison with HRRRv4 for the period 2020–21. The changes implemented with HRRRv2 were not particularly targeted at improved ceiling forecasts, but the impacts of changes are seen in the differences between the blue and green curves in Fig. 18. HRRRv2 exhibits an increased high bias in occurrence of low clouds for most thresholds and forecast lengths, as well as a reduction in CSI for 1000-ft ceilings and an improvement in CSI for 500-ft ceilings. HRRRv3 featured several changes aimed at improving low cloud forecasts, including consistent cloud building from both satellite and METAR cloud observations below 1200 m AGL, and a decrease in the assumed cloud water or
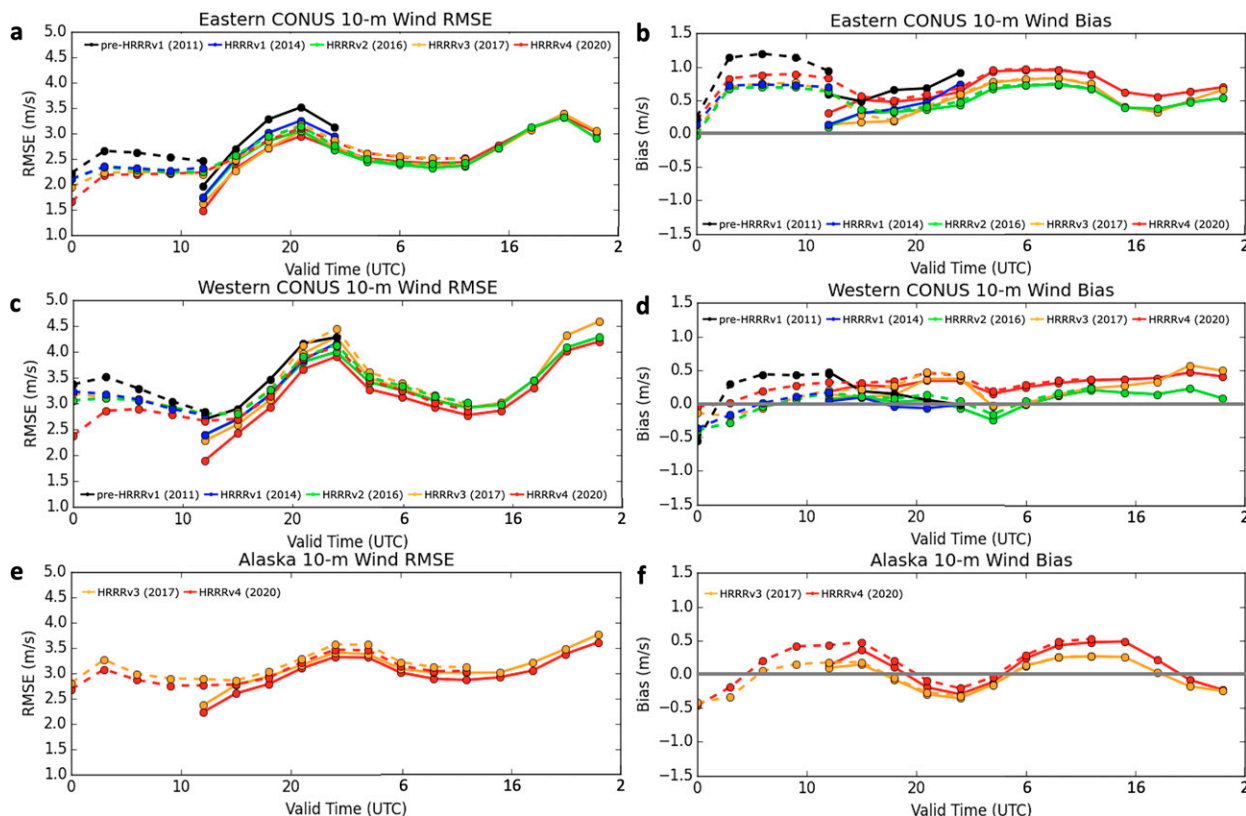
FIG. 16. As in Fig. 10, but for 10-m wind speed.

cloud ice mixing ratio when clouds are built (implying smaller cloud hydrometeors and longer cloud retention). Physics changes to improve the representation of cloud cover within the MYNN PBL scheme also contribute to cloud improvements in HRRRv3. These changes led to an improvement in CSI for occurrence of ceilings at 3000, 1000, and 500 ft (Fig. 18, orange curves). In addition, a high frequency bias in occurrence of low clouds is generally reduced (improved; Fig. 18). HRRRv4 features substantial physics changes for low-cloud retention, as well as storm-scale covariance information for DA. This leads to a reduction in the magnitude of the frequency bias drop across lead times for all ceiling thresholds (cf. orange and red curves in Fig. 18). For 3000- and 500-ft ceilings, the HRRRv4 exhibits a near neutral frequency bias at forecast lengths beyond 1 h (Fig. 18).

Figure 19 shows surface visibility forecasts for the operational HRRR, evaluated for six month periods to allow inclusion of HRRRv1 (ceiling results are not available prior to January 2016). Surface visibility is diagnosed in HRRR as described by Benjamin et al. (2020). The diagnosis is dependent upon low-level relative humidity and precipitation hydrometeors, and, with HRRRv4, also upon smoke concentrations. A dramatic improvement in CSI is evident in HRRRv2 (Fig. 19, green curves) as compared with HRRRv1 (blue curves). HRRRv3 visibility forecasts generally exhibited a higher CSI than HRRRv2 forecasts for the shorter forecast lengths (Fig. 19, green to orange curves). This improvement in CSI was associated with a reduced frequency bias for 1-mi visibility (Fig. 19, orange

curve with stars). Storm-scale covariance information used in the HRRRv4 DA led to a reduced drop in CSI for the first few hours of the forecast compared to HRRRv3 (Fig. 19; red curves).

Improved cloud ceiling and surface visibility representation is important for many applications, especially surface and aviation transportation. In this section, we have summarized HRRR performance for these variables across a number of important thresholds, demonstrating the forecast improvements stemming from both DA and model physics development. More details on the stratiform cloud hydrometeor analysis system, which is a critical component of the HRRR's overall cloud ceiling and visibility forecast capabilities, are provided by Benjamin et al. (2021), with quantification of the forecast impacts in the RAP.

5. Future directions and conclusions

The HRRR represents the first operational hourly updating CAM in the United States, and as such, it has proven to be a critical tool for forecast users interested in short-range, high-impact weather forecasting. The HRRR adds increased capability and value over the output from the mesoscale RAP system, explicitly forecasting the evolution of convective storms as well as orographic effects on scales of tens of kilometers. The benefits of convection-allowing grid spacing (~3–4 km) over mesoscale grid spacing (~10 km) have been documented for convective storms by Done et al. (2004) and Weisman et al. (2008). While
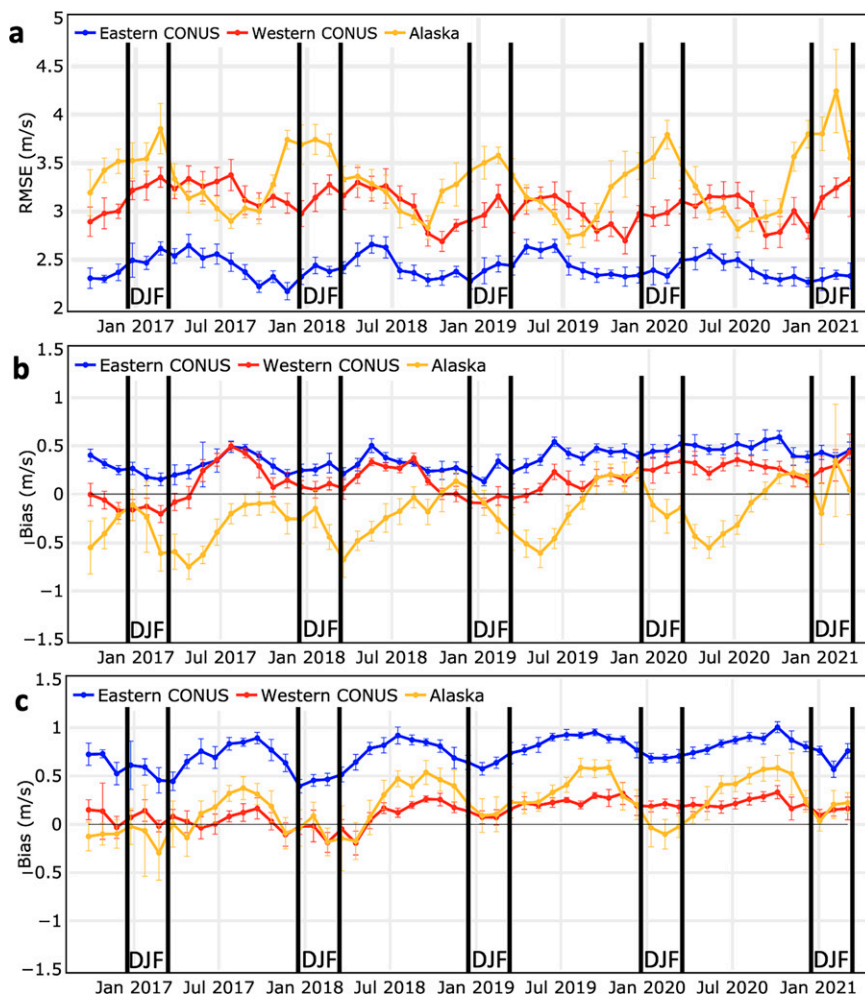
FIG. 17. As in Fig. 11, but for 10-m wind speed.

major progress has been achieved since the implementation of HRRRv1, many forecasting challenges persist for future generations of CAM guidance to address. These challenges include, for example, convective initiation, MCS evolution, PBL development in complex terrain, and the initialization of clouds and precipitating systems in general. The HRRR represents a baseline against which future CAM forecasts can be evaluated, although further work is needed to define a statistical baseline for evaluating forecast improvements at longer lead times, and across broader geographic domains.

The fourth version of the HRRR, HRRRv4, is the last operational version of the HRRR. Beyond the HRRR, NOAA is moving toward the Unified Forecasting System (UFS), based on the Finite Volume cubed sphere (FV3; Chen et al. 2013), to consolidate NWP development within the United States; this effort will involve wide collaboration with many laboratories and the university community. Within this framework, CAM DA and model physics development which has focused on the HRRR system has shifted toward the development of a UFS-based Rapid Refresh Forecast System (RRFS), with a view toward replacing the HRRR system later this decade.

Such an RRFS system will enable developers from various laboratories and universities to collaborate on common problems, and advance the state of the science for CAM NWP.

Future development efforts within the RRFS era will involve an increased focus on ensemble design, which will pave the way for an explicit representation of uncertainty in forecasting, and will also lead to major benefits in storm-scale DA through more realistic covariance structures. The HRRR Ensemble system (HRRRE; Dowell et al. 2016), which ran experimentally during much of the recent HRRR development era, was an initial step in this direction. Ongoing development is focusing on a prototype RRFS ensemble (RRFSE), experimenting with both initial condition perturbations as well as physics perturbations (Kalina et al. 2021). The RRFSE is a prototype single-core ensemble system with 9 members, and is being evaluated for potential to replace the operational ensemble HREF system (Roberts et al. 2019).

The RRFS era could also feature an increasing linkage with important Earth system components. The HRRRv4 has taken small steps in this direction with the usage of one-way coupling for lake surface temperatures and ice cover over the Great Lakes, as well as the introduction of a smoke tracer
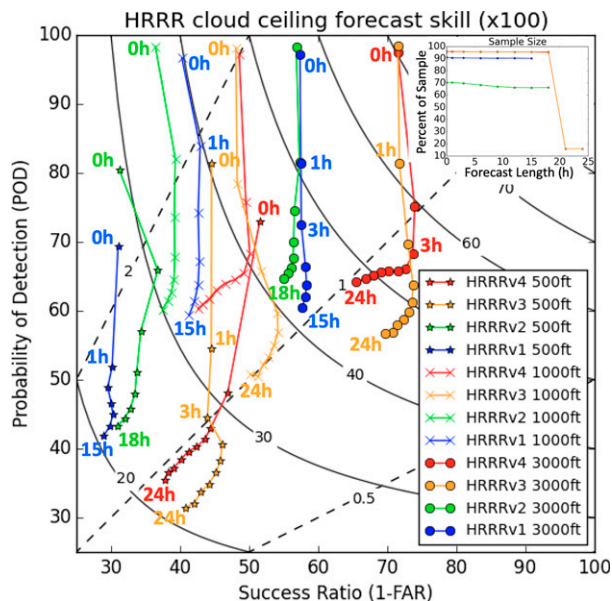
FIG. 18. Performance diagram for HRRR cloud ceiling forecasts for CONUS, verified against METAR observations (1700–1800 stations per hour). The x axis is success ratio (1 − false alarm ratio), and the y axis is probability of detection. HRRR results for HRRRv1 (December 2014–November 2015), HRRRv2 (December 2016–November 2017), HRRRv3 (December 2018–November 2019), and HRRRv4 (December 2020–November 2021). Results are shown for 3000-, 1000-, and 500-ft ceiling occurrence, for 0-, 1-, 3-, 6-, 9-, 12-, 15-, 18-, 21-, and 24-h forecasts for threshold. Note that HRRRv1 results only extend through 15 h, and HRRRv2 results only extend through 18 h.
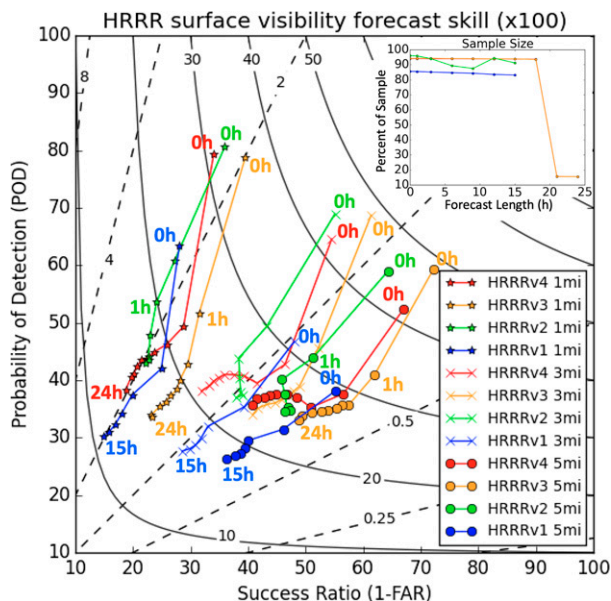
FIG. 19. Performance diagram for HRRR surface visibility forecasts for CONUS, verified against METAR observations (1700–1800 stations per hour). The x axis is success ratio (1 − false alarm ratio), and the y axis is probability of detection. HRRR results for HRRRv1 (January–July 2016), HRRRv2 (January–July 2017), HRRRv3 (January–July 2019), and HRRRv4 (January–July 2021). Results are shown for 5-, 3-, and 1-mi visibility occurrence, for 0-, 1-, 3-, 6-, 9-, 12-, 15-, 18-, 21-, and 24-h forecasts for each threshold. Note that HRRRv1 and HRRRv2 results only extend through 15 h.

with radiation interactions. Future Earth system components to be increasingly coupled with meteorology include the ocean, sea ice, land surface vegetation, additional aerosols including blowing dust and volcanic ash, and chemical species for air quality forecasting. Including the effects of these complicated systems in a computationally efficient manner within a unified NWP system is a challenging goal that will require broad collaboration within and beyond NOAA, but will lead to significant forecast improvements.

As computational resources increase in the coming years, the horizontal extent of the modeling domains used for CAM systems will expand in order to improve treatment of incoming weather systems and account for long-range transport of chemical species. Increasing resources will reduce the need for nested domains, although very high-resolution CAM configurations will remain important for local events and regions of interest (e.g., Mailhot et al. 2012; Golding et al. 2014). Expanding CAM domains will necessitate the development of a three-dimensional global radar analysis for use in initializing precipitating systems, as well as improved storm-scale satellite DA, which represent major technical challenges in their own right. However, enabling the development of a global rapidly updating CAM system will benefit many aspects of society across our increasingly connected world, and save lives and property in many regions subject to extreme weather.

(https://console.cloud.google.com/marketplace/product/noaa-public/hrrr?project=python-232920&pli=1). While experimental HRRR output in this study is not publicly available, the operational HRRR output available from the cloud providers could be used to evaluate the results of this study. Real-time hourly forecasts are available from the NOAA/National Centers for Environmental Prediction (NCEP) Central Operations (NCO) (https://nomads.ncep.noaa.gov/pub/data/nccf/com/hrrr/prod/).

REFERENCES

Augustine, J. A., J. Deluisi, and C. N. Long, 2000: SURFRAD—A national surface radiation budget network for atmospheric research. *Bull. Amer. Meteor. Soc.*, **81**, 2341–2357, https://doi.org/10.1175/1520-0477(2000)081<2341:SANSRB>2.3.CO;2.

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

——, E. P. James, J. M. Brown, E. J. Szoke, J. S. Kenyon, and R. Ahmadov, 2020: Diagnostic fields for hourly updated NOAA weather models. NOAA Tech Memo. OAR GSL, 66, 54 pp., https://repository.library.noaa.gov/view/noaa/24212.

——, and Coauthors, 2021: Stratiform cloud hydrometeor assimilation for HRRR and RAP model short-range weather prediction. *Mon. Wea. Rev.*, **149**, 2673–2694, https://doi.org/10.1175/MWR-D-20-0319.1.

Blaylock, B. K., and J. D. Horel, 2020: Comparison of lightning forecasts from the High-Resolution Rapid Refresh model to geostationary lightning mapper observations. *Wea. Forecasting*, **35**, 401–416, https://doi.org/10.1175/WAF-D-19-0141.1.

Bytheway, J. L., and C. D. Kummerow, 2015: Toward an object-based assessment of high-resolution forecasts of long-lived convective precipitation in the central US. *J. Adv. Model. Earth Syst.*, **7**, 1248–1264, https://doi.org/10.1002/2015MS000497.

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

Chen, X., N. Andronova, B. van Leer, J. Penner, J. P. Boyd, C. Jablonowski, and S.-J. Lin, 2013: A control-volume model of the compressible Euler equations with a vertical Lagrangian coordinate. *Mon. Wea. Rev.*, **141**, 2526–2544, https://doi.org/10.1175/MWR-D-12-00129.1.

Chow, F. K., and Coauthors, 2022: High-resolution smoke forecasting for the 2018 Camp Fire in California. *Bull. Amer. Meteor. Soc.*, **103**, E1531–E1552, https://doi.org/10.1175/BAMS-D-20-0329.1.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the weather research and forecasting (WRF) model. *Atmos. Sci. Lett.*, **5**, 110–117, https://doi.org/10.1002/asl.72.

Dougherty, K. J., J. D. Horel, and J. E. Nachamkin, 2021: Forecast skill for California heavy precipitation periods from the High-Resolution Rapid Refresh model and the Coupled Ocean-Atmosphere Mesoscale Prediction System. *Wea. Forecasting*, **36**, 2275–2288, https://doi.org/10.1175/WAF-D-20-0182.1.

Dowell, D. C., and Coauthors, 2016: Development of a High-Resolution Rapid Refresh Ensemble (HRRRE) for severe weather forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 8B.2, https://ams.confex.com/ams/28SLS/webprogram/Paper301555.html.

——, and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Wea. Forecasting*, in press, https://doi.org/10.1175/WAF-D-21-0151.1.

Duda, J. D., and D. D. Turner, 2021: Large-sample application of radar reflectivity object-based verification to evaluate HRRR warm-season forecasts. *Wea. Forecasting*, **36**, 805–821, https://doi.org/10.1175/WAF-D-20-0203.1.

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. *Wea. Forecasting*, **36**, 1925–1944, https://doi.org/10.1175/WAF-D-21-0081.1.

Fovell, R. G., and A. Gallagher, 2020: Boundary layer and surface verification of the High-Resolution Rapid Refresh, version 3. *Wea. Forecasting*, **35**, 2255–2278, https://doi.org/10.1175/WAF-D-20-0101.1.

Golding, B. W., and Coauthors, 2014: Forecasting capabilities for the London 2012 Olympics. *Bull. Amer. Meteor. Soc.*, **95**, 883–896, https://doi.org/10.1175/BAMS-D-13-00102.1.

Griffin, S. M., J. A. Otkin, C. M. Rozoff, J. M. Sieglaff, L. M. Cronce, and C. R. Alexander, 2017a: Methods for comparing simulated and observed satellite infrared brightness temperatures and what do they tell us? *Wea. Forecasting*, **32**, 5–25, https://doi.org/10.1175/WAF-D-16-0098.1.

——, ——, ——, ——, ——, ——, T. L. Jensen, and J. K. Wolff, 2017b: Seasonal analysis of cloud objects in the High-Resolution Rapid Refresh (HRRR) model using object-based verification. *J. Appl. Meteor. Climatol.*, **56**, 2317–2334, https://doi.org/10.1175/JAMC-D-17-0004.1.

Hu, M., S. G. Benjamin, T. T. Ladwig, D. C. Dowell, S. S. Weygandt, C. R. Alexander, and J. S. Whitaker, 2017: GSI three-dimensional ensemble-variational hybrid data assimilation using a global ensemble for the regional Rapid Refresh model. *Mon. Wea. Rev.*, **145**, 4205–4225, https://doi.org/10.1175/MWR-D-16-0418.1.

Ikeda, K., M. Steiner, J. Pinto, and C. Alexander, 2013: Evaluation of cold-season precipitation forecasts generated by the hourly updating High-Resolution Rapid Refresh model. *Wea. Forecasting*, **28**, 921–939, https://doi.org/10.1175/WAF-D-12-00085.1.

——, ——, and G. Thompson, 2017: Examination of mixed-phase precipitation forecasts from the High-Resolution Rapid Refresh model using surface observations and sounding data. *Wea. Forecasting*, **32**, 949–967, https://doi.org/10.1175/WAF-D-16-0171.1.

Jirak, I. L., A. J. Clark, B. Roberts, B. T. Gallo, and S. J. Weiss, 2018: Exploring the optimal configuration of the High Resolution Ensemble Forecast System. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 14B.6, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html.

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, https://doi.org/10.1175/WAF-D-20-0098.1.

Lee, T. R., M. Buban, D. D. Turner, T. P. Meyers, and C. B. Baker, 2019: Evaluation of the High-Resolution Rapid Refresh (HRRR) model using near-surface meteorological and flux observations from northern Alabama. *Wea. Forecasting*, **34**, 635–663, https://doi.org/10.1175/WAF-D-18-0184.1.

Lundquist, J., M. Hughes, E. Gutmann, and S. Kapnick, 2019: Our skill in modeling mountain rain and snow is bypassing the skill

of our observational networks. *Bull. Amer. Meteor. Soc.*, **100**, 2473–2490, https://doi.org/10.1175/BAMS-D-19-0001.1.

Mailhot, J., J. A. Milbrandt, A. Giguère, R. McTaggart-Cowan, A. Erfani, B. Denis, A. Glazer, and M. Vallée, 2012: An experimental high-resolution forecast system during the Vancouver 2010 Winter Olympic and Paralympic Games. *Pure Appl. Geophys.*, **171**, 209–229, https://doi.org/10.1007/s00024-012-0520-6.

McCorkle, T. A., J. D. Horel, A. A. Jacques, and T. Alcott, 2018: Evaluating the experimental High-Resolution Rapid Refresh-Alaska modeling system using USArray pressure observations. *Wea. Forecasting*, **33**, 933–953, https://doi.org/10.1175/WAF-D-17-0155.1.

Mittermaier, M. P., 2014: A strategy for verifying near-convection-resolving model forecasts at observing sites. *Wea. Forecasting*, **29**, 185–204, https://doi.org/10.1175/WAF-D-12-00075.1.

Pichugina, Y. L., and Coauthors, 2019: Spatial variability of winds and HRRR–NCEP model error statistics at three Doppler-lidar sites in the wind-energy generation region of the Columbia River Gorge. *J. Appl. Meteor. Climatol.*, **58**, 1633–1656, https://doi.org/10.1175/JAMC-D-18-0244.1.

Pinto, J. O., J. A. Grim, and M. Steiner, 2015: Assessment of the High-Resolution Rapid Refresh model's ability to predict mesoscale convective systems using object-based evaluation. *Wea. Forecasting*, **30**, 892–913, https://doi.org/10.1175/WAF-D-14-00118.1.

Radford, J. T., G. M. Lackmann, and M. A. Baxter, 2019: An evaluation of snowband predictability in the High-Resolution Rapid Refresh. *Wea. Forecasting*, **34**, 1477–1494, https://doi.org/10.1175/WAF-D-19-0089.1.

Randriamampianina, R., N. Bormann, M. Koltzow, H. Lawrence, I. Sandu, and Z. Q. Wang, 2021: Relative impact of observations on a regional Arctic numerical weather prediction system. *Quart. J. Roy. Meteor. Soc.*, **147**, 2212–2232, https://doi.org/10.1002/qj.4018.

Roberts, B. B., T. Gallo, I. L. Jirak, and A. J. Clark, 2019: The High Resolution Ensemble Forecast (HREF) system: Applications and performance for forecasting convective storms. *2019 Fall Meeting*, San Francisco, CA, Amer. Geophys. Union, Abstract A310-2797, https://doi.org/10.1002/essoar.10501462.1.

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

Smith, T. M., and Coauthors, 2016: Multi-radar multi-sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

——, G. S. Romine, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection-allowing model. *Wea. Forecasting*, **35**, 1981–2000, https://doi.org/10.1175/WAF-D-20-0036.1.

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

Tustison, B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res.*, **106**, 11 775–11 784, https://doi.org/10.1029/2001JD900066.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, https://doi.org/10.1175/2007WAF2007005.1.

Weygandt, S. S., S. G. Benjamin, M. Hu, C. R. Alexander, T. G. Smirnova, and E. P. James, 2022: Radar reflectivity-based model initialization using specified latent heating (Radar-LHI) within a diabatic digital filter or pre-forecast integration. *Wea. Forecasting*, in press, https://doi.org/10.1175/WAF-D-21-0142.1.

Wilczak, J. M., and Coauthors, 2019: The Second Wind Forecast Improvement Project (WFIP2): Observational field campaign. *Bull. Amer. Meteor. Soc.*, **100**, 1701–1723, https://doi.org/10.1175/BAMS-D-18-0035.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Zhang, J., and Coauthors, 2016: Multi-radar multi-sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–637, https://doi.org/10.1175/BAMS-D-14-00174.1.