Evaluation of the Near-Surface Variables in the HRRR Weather Model Using Observations from the ARM SGP Site

SIWEI HE,[a,b] DAVID D. TURNER,[b] STANLEY G. BENJAMIN,[a,b] JOSEPH B. OLSON,[b] TATIANA G. SMIRNOVA,[a,b] AND TILDEN MEYERS[c]

[a] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[b] *NOAA/Global Systems Laboratory, Boulder, Colorado*
[c] *NOAA/Air Resources Laboratory, Oak Ridge, Tennessee*

(Manuscript received 5 January 2023, in final form 31 March 2023, accepted 12 April 2023)

ABSTRACT: The performance of version 4 of the NOAA High-Resolution Rapid Refresh (HRRR) numerical weather prediction model for near-surface variables, including wind, humidity, temperature, surface latent and sensible fluxes, and longwave and shortwave radiative fluxes, is examined over the Atmospheric Radiation Measurement (ARM) Southern Great Plains (SGP) region. The study evaluated the model's bias and bias-corrected mean absolute error relative to the observations on different time scales. Forecasts of near-surface geophysical variables at five SGP sites (HRRR at 3-km scale) were found to agree well with observations, but some consistent observation–forecast differences also occurred. Sensible and latent heat fluxes are the most challenging variables to be reproduced. The diurnal cycle is the main temporal scale affecting observation–forecast differences of the near-surface variables, and almost all of the variables showed different biases throughout the diurnal cycle. Results show that the overestimation of downward shortwave and the underestimation of downward longwave radiative flux are the two major biases found in this study. The timing and magnitude of downward longwave flux, wind speed, and sensible and latent heat fluxes are also different with contributions from model representations, data assimilation limitations, and differences in scales between HRRR and SGP sites. The positive bias in downward shortwave and negative bias in longwave radiation suggests that the model is underestimating cloud fraction in the study domain. The study concludes by showing a brief comparison with version 3 of the HRRR and shows that version 4 has better performance in almost all near-surface variables.

SIGNIFICANCE STATEMENT: A correct representation of the near-surface variables is important for numerical weather prediction models. This study investigates the capability of the latest NOAA High-Resolution Rapid Refresh (HRRRv4) model in simulating the near-surface variables by comparing against the Atmospheric Radiation Measurement (ARM) Southern Great Plains (SGP) in situ observations. Among others, we find that the surface heat fluxes, such as sensible and latent heat fluxes, are the most difficult variables to be reproduced. This study also shows that the diurnal cycle has the dominant impact on the model's performance, which means the majority of the outputted near-surface variables have the strong diurnal cycle in their bias errors.

KEYWORDS: Hydrometeorology; Surface fluxes; Numerical weather prediction/forecasting

1. Introduction

Weather information is a key component of environmental situational awareness, and the 3-km High-Resolution Rapid Refresh (HRRR) model provides a critical component of NOAA's suite of weather guidance models focusing on situational awareness and short-range forecasting (Benjamin et al. 2016). Version 1 of the HRRR (HRRRv1) was implemented at the National Centers for Environmental Prediction (NCEP) in September 2014. Development has continued, with advances in both model configuration (e.g., physics parameterizations, diffusion) and data assimilation over the years, resulting in improved models, with HRRRv2 being implemented at NCEP in August 2016, HRRRv3 in July 2018, and HRRRv4 in December 2020. Before the implementation of HRRRv4, the modeling system had been run in test mode since August 2019, providing over 15 months of data for analysis and comparison with HRRRv3. A full description of the HRRR, including background on the different versions, is provided by Dowell et al. (2022, hereinafter D22).

There are several published studies evaluating the performance of different versions of the HRRR, but many of them focused on the precipitation forecast (e.g., Yue and Gebremichael 2020; Bytheway et al. 2017; Cai and Dumais 2015; Gowan et al. 2018; Ikeda et al. 2017; Duda and Turner 2021; English et al. 2021). For example, Yue and Gebremichael (2020) evaluated the accuracy of short-range forecasts from the HRRRv2 model for five extreme storms in the United States. The study concluded that the HRRRv2 model provides relatively reliable forecasts, but the forecasts were mostly biased, overestimating rainfall in hurricanes but underestimating precipitation during tropical storms that move into the center of the country. Gowan et al. (2018) evaluated the cool-season precipitation of HRRRv2 over the

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JAMC-D-23-0003.s1.

Corresponding author: Siwei He, siwei.he@noaa.gov

DOI: 10.1175/JAMC-D-23-0003.1

western United States by comparing it against Snowpack Telemetry (SNOTEL) observations and Parameter-Elevation Regressions on Independent Slopes Model (PRISM) analyses. They found that the HRRRv2 can adequately simulate orographic precipitation with good performance compared with other mesoscale models. English et al. (2021) demonstrated that HRRRv4 was able to accurately predict precipitation over central California in atmospheric river events, although there is a slight underestimate in the Central Valley and overestimate in the Sierra Nevada. Cai and Dumais (2015) evaluated the experimental version of the HRRR model's precipitation forecasts during the summer of 2010 over the eastern two-thirds of the contiguous United States. It was found that the HRRR model was able to forecast convective storm characteristics rather well, either as a function of time of day or as a function of storm size, although the model tended to underestimate the total number and total area of convective storms. Ikeda et al. (2017) examined the HRRRv2 model's ability to forecast the surface precipitation phase and concluded that the HRRRv2 model was able to represent the overall vertical thermodynamic structure in the mixed-phase precipitation regions. So far, only a few studies have examined other near-surface variables (e.g., Lee et al. 2019; Fovell and Gallagher 2020, 2022; Wagner et al. 2019). For example, Lee et al. (2019) evaluated the HRRRv2 model's capability in simulating the near-surface meteorological fields and surface energy balance at two locations in northern Alabama. They found that the HRRRv2 model accurately simulated the observations of near-surface air and dewpoint temperature, but the model did not simulate well the observed sensible, latent, and ground heat fluxes. Wagner et al. (2019) used temperature and humidity profiles retrieved from the Atmospheric Emitted Radiance Interferometer (AERI; Turner and Blumberg 2019) radiance measurements to calculate convective available potential energy (CAPE) and compared these values with output from the HRRRv2. They showed that HRRRv2 diurnal distribution of CAPE lagged by 2 to 4 h compared with the AERI observations. James et al. (2022, hereinafter J22) performed a more comprehensive evaluation of the HRRR model (versions 1–4) using Meteorological Aerodrome Report (METAR) surface observations, radar, and surface radiation data over the CONUS domain.

These parameterizations for land surface, surface layer, and planetary boundary layer schemes have been continuously improved over the lifetime of the HRRR model development (e.g., Smirnova et al. 2016; Olson et al. 2019a; Angevine et al. 2020; D22). This study will comprehensively evaluate the performance of HRRRv4 in simulating 10-m wind, 2-m air temperature and humidity, surface radiation, and surface turbulent heat fluxes for a particularly well-observed substate area in north-central Oklahoma. This study will also make comparisons of HRRRv4 and HRRRv3. The goal of this work is to provide guidance for further improving the HRRRv4's capability in simulating near-surface variables. In contrast to J22, the purposes of this study are to 1) present a detailed evaluation of HRRR performance in simulating near-surface variables over a particularly well-observed region in Oklahoma using in situ observations; 2) reveal features of model performance over the southern Great Plains (SGP) region that could be obscured by averaging over large areas; and 3) quantify the forecast performance of HRRRv4 over different time scales. The SGP region was identified as one of the regions with strong land–atmosphere coupling (Koster et al. 2004), and we hope this study provides useful information on the sensitivities and feedback mechanisms for studying land–atmosphere interactions using HRRRv4.

2. Data and method

a. Data

1) ARM SGP

The SGP atmospheric observatory was the first field measurement site established by the Atmospheric Radiation Measurement (ARM) program (Turner and Ellingson 2016). The SGP observatory consists of in situ and remote sensing instrument clusters (e.g., Sisterson et al. 2016; Yang et al. 2006). Unique aspects at the SGP observatory are the multiple locations where temperature, humidity, wind, and turbulence profiles are observed, as well as both collocated and distributed surface energy-budget stations. Land cover of these sites and their surrounding areas is given in Fig. 1. In the HRRR model, the vegetation type of these grids with sites is croplands or grasslands. Measurements from the E13 (collocated with the Central Facility, which is southeast of Lamont, Oklahoma), E32, E37, E39, and E41 (Fig. 1) are used for the analysis in this work. These stations cover a 70-km (north–south) by 80-km (east–west) area.

The measurements of the following variables of the SGP sites are used in this study: 2-m air temperature (Temp), 2-m water vapor mixing ratio (WVMR), 10-m wind speed (Wspd), surface sensible heat flux (Sflux), surface latent heat flux (Lflux), surface upward longwave radiation (LWup), surface downward longwave radiation (LWdn), surface upward shortwave radiation (SWup), and surface downward shortwave radiation (SWdn). Among these variables, 2-m air temperature and 10-m wind speed are from the surface meteorological systems (https://doi.org/10.5439/1025220), sensible heat flux and latent heat fluxes are from the energy balance Bowen ratio (https://doi.org/10.5439/1023895, at site E32) and the eddy correlation flux measurement systems (https://doi.org/10.5439/1097546, at sites E13, E37, E39, and E41), the shortwave and longwave radiative fluxes are from solar infrared radiation stations (https://doi.org/10.5439/1025277), and the water vapor mixing ratio is derived from the measurements of the surface meteorology systems. The observed data of the ARM observatory can be downloaded from its official site (https://www.arm.gov/). Because of a system issue in the observations, latent heat flux measurements from October 2019 to June 2020 are not usable at these sites.

2) HRRR FORECAST MODEL

The HRRR is an operational 3-km-resolution weather prediction system and is an hourly updated, cloud-resolving, convection-allowing atmospheric model (Benjamin et al. 2016; D22) with hourly data assimilation. It has been running operationally at NOAA/NWS/NCEP since 2014. The HRRR

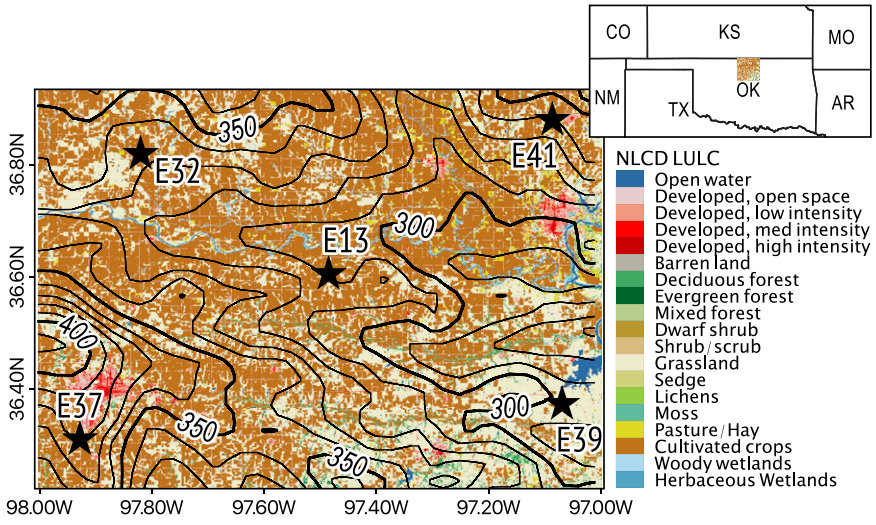

FIG. 1. The five ARM SGP sites of the study domain (stars). The solid black lines represent elevation (above sea level) contours (m), and land cover is from the National Land Cover Database (NLCD; Homer et al. 2015).

has been undergoing continuous improvements, which are described online (https://rapidrefresh.noaa.gov/hrrr/).

HRRRv4 is the latest version (as of this writing) and became operational at NCEP in December 2020. As part of the evaluation process used by the NWS, the HRRRv4 code was run in an operational-like manner within the NOAA Global Systems Laboratory from August 2019 to December 2020. The processes and fluxes near the surface treated in the HRRR are represented in Fig. 1 in Smirnova et al. (2000). HRRRv4 is the primary version assessed in this study, but its predecessor, HRRRv3, was also used for comparison. HRRRv4 uses the Rapid Update Cycle (RUC) land surface model (LSM) (Smirnova et al. 2016; He et al. 2021, see their Table A1 for more details) for computing surface fluxes. It also uses the Mellor–Yamada–Nakanishi–Niino (MYNN) eddy-diffusivity/mass-flux (EDMF) scheme (Olson et al. 2019a, 2021) for computing turbulent exchanges of moisture, heat, and momentum in the atmosphere. HRRRv4 computes the subgrid-scale cloud properties via MYNN-EDMF, which is coupled with the Rapid Radiative Transfer Model for GCMs (RRTMG, Iacono et al. 2008) for computing shortwave and longwave radiation. HRRRv4 uses MODIS albedo climatology and the real-time VIIRS greenness fraction. HRRRv4 also introduced a detailed treatment of larger and small lakes to improve lake–atmosphere fluxes (Benjamin et al. 2022b). One of the key features of HRRRv4 is that its representation of both stratiform and convective subgrid-scale boundary layer clouds has been improved (Olson et al. 2019b; D22), and the previous large positive bias of incoming shortwave radiation has been reduced. This led to the improvements of HRRRv4 over the CONUS domain in general (J22), and here we examine the impacts of these improvements over the smaller SGP domain. The forecast lead times of HRRRv4 range from 1 to 48 h. Hourly data assimilation to initialize the HRRR includes use of surface METAR, aircraft, WSR-88D, and satellite and ceilometer cloud data (D22; Benjamin et al. 2016, 2021). HRRR data assimilation also extends to soil temperature and moisture (Benjamin et al. 2022a).

In HRRRv4, 10-m wind speed is a diagnosed variable calculated from the wind speed at the middle of the first model layer using a neutral-log interpolation (Olson et al. 2021). The 2-m temperature (water vapor mixing ratio) is also a diagnosed variable interpolated from the surface potential temperature (surface mixing ratio) and the potential temperature (mixing ratio) at the first model level weighting by ratio of the thermal (moisture) resistances (Olson et al. 2021). The surface latent heat flux is calculated from the total evapotranspiration, and the sensible heat flux is calculated using the heat transfer coefficient from the MYNN surface-layer scheme and the temperature difference between the first model level and the skin level (Smirnova et al. 2016). The surface heat fluxes are estimated in the RUC LSM.

In this study, we have used HRRR output data from a 15-month period from August 2019 to October 2020 for evaluating the model's performance. In the analysis, our investigation mainly uses the 6-h forecast from the HRRRv4 for comparison with the observations from the five ARM SGP sites. The forecasts from 1 to 48 h were used in studying the effects of lead times on the model's performance. HRRR output variables are instantaneous, valid at a specific 20-s time step, and without any temporal smoothing.

b. *Method to compare model with observations*

In this study, we use bias error and the bias-corrected mean absolute error (BCMAE) of these variables to characterize the performance of the model. They are calculated as

$$\text{bias} = \frac{1}{n}\sum_{j=1}^{n}(x_{f,j} - x_{o,j}) \quad \text{and} \tag{1}$$

$$\text{BCMAE} = \frac{1}{n}\sum_{j=1}^{n}|x_{f,j} - x_{o,j} - \text{bias}|, \tag{2}$$

where $x_{o,j}$ is the jth observed value, $x_{f,j}$ is the jth forecast value, and n is the number of pairs in the comparison. The BCMAE is particularly useful to evaluate the improvement of model skill; while it is easy to change a bias within a model, it is hard to reduce the BCMAE without improving the model's ability to represent the physical evolution of the atmosphere (e.g., Turner et al. 2004). For the comparison between the models and the observations, data from the model grid cells around the observations were obtained at hourly resolution and then interpolated using a bilinear method.

Using these two statistical metrics, we evaluated the performance of HRRRv4 against observations in multiple ways, including 1) monthly performance of HRRRv4 over the 15-month time period, 2) total performance of HRRRv4 in the cold (December 2019–February 2020) and warm (June–August 2020) seasons, 3) diurnal performance of HRRRv4 in the same cold and warm seasons, and 4) performance of the HRRRv4 as a function of forecast lead time in the warm season. We also compared the performance of HRRRv3 and HRRRv4 in the cold and warm seasons, respectively, to investigate the changes in performance between the two model versions.

3. Results

As shown in Fig. 1, we used data from five ARM sites in the SGP domain. The soil properties are either loam, silt loam, or clay loam, and the primary land use is either croplands or grasslands. An analysis (not shown) demonstrates that the near-surface variables predicted by the HRRR are generally consistent across these five sites; thus, we have combined the data from the five locations to improve the sampling in the subsequent statistics. However, a brief discussion on the variation of the model's performance among these five sites is given in the online supplemental material for readers interested in this topic.

a. Monthly performance

Near-surface variables can have a strong annual cycle. To study the monthly variation of the model's performance, monthly statistical metrics of the 15-month (August 2019–October 2020) 6-h forecasts were calculated. As near-surface variables are heavily impacted by the diurnal cycle, daytime and nighttime statistical metrics were calculated separately. Also, to avoid the morning and afternoon transition effects, daytime and nighttime metrics were computed from 1000 to 1500 and from 2200 to 0300 local time (LT; i.e., central daylight time), respectively. Since the model's performance over these five sites is largely consistent, the same metrics for each variable at the five sites were combined and only their mean and variation (one sigma) were presented. Figure 2 shows bias errors separated by day and night of the different months for the eight variables; Fig. 3 shows the same, but for the BCMAEs of these variables.

The monthly bias errors (Fig. 2) of different variables show that almost all variables have variations over different months, and the bias errors were generally larger in the daytime than in the nighttime. The 2-m air temperature was overestimated by about 1 K for all months in the nighttime, underestimated in the daytime over September, October, and November of the 15 months, but close to zero for other months (more results about the diurnal variation are in the later section) (Fig. 2a). For both nighttime and daytime, HRRRv4 underestimated 2-m water vapor mixing ratio during the warm months (Fig. 2b, close to zero for other months) and also underestimated upward (Fig. 2c) and downward (Fig. 2d) longwave radiation during most of this analysis period, which was especially pronounced in the daytime. The model always overestimated downward shortwave radiation (Fig. 2e). The bias errors of sensible and latent heat fluxes had the opposite signs (Figs. 2g,h). When the sensible heat flux was underestimated in the daytime in August, September, and October 2019, latent heat flux was overestimated, and when the

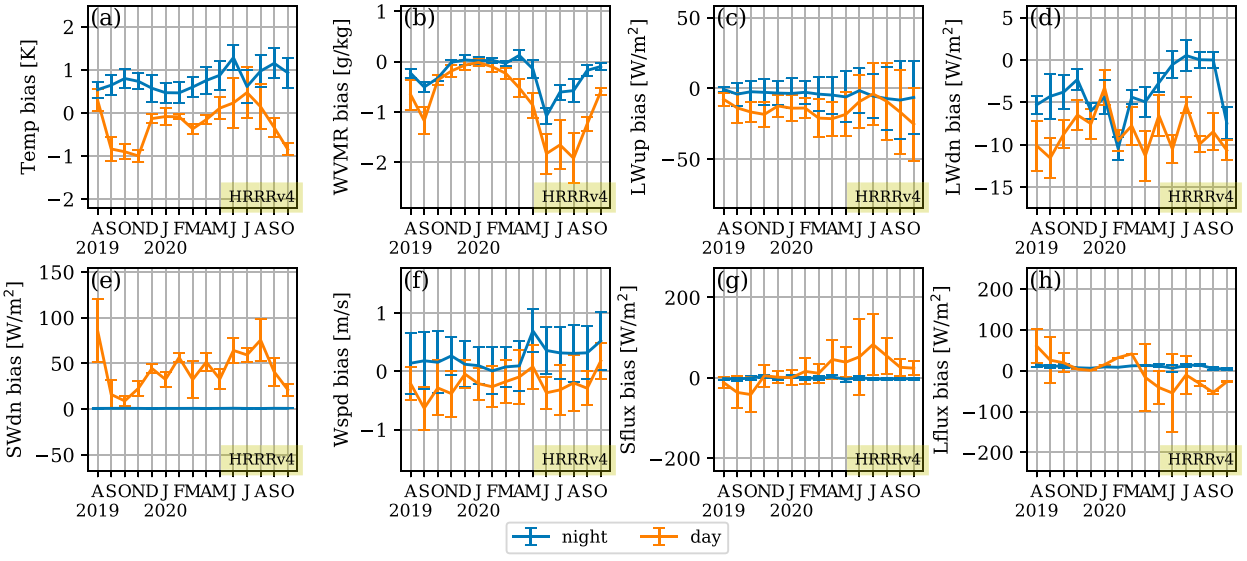

FIG. 2. Monthly daytime (1000–1500 LT; orange) and nighttime (2200–0300 LT; blue) bias of the eight different variables between the 6-h HRRRv4 forecasts and observed variables.

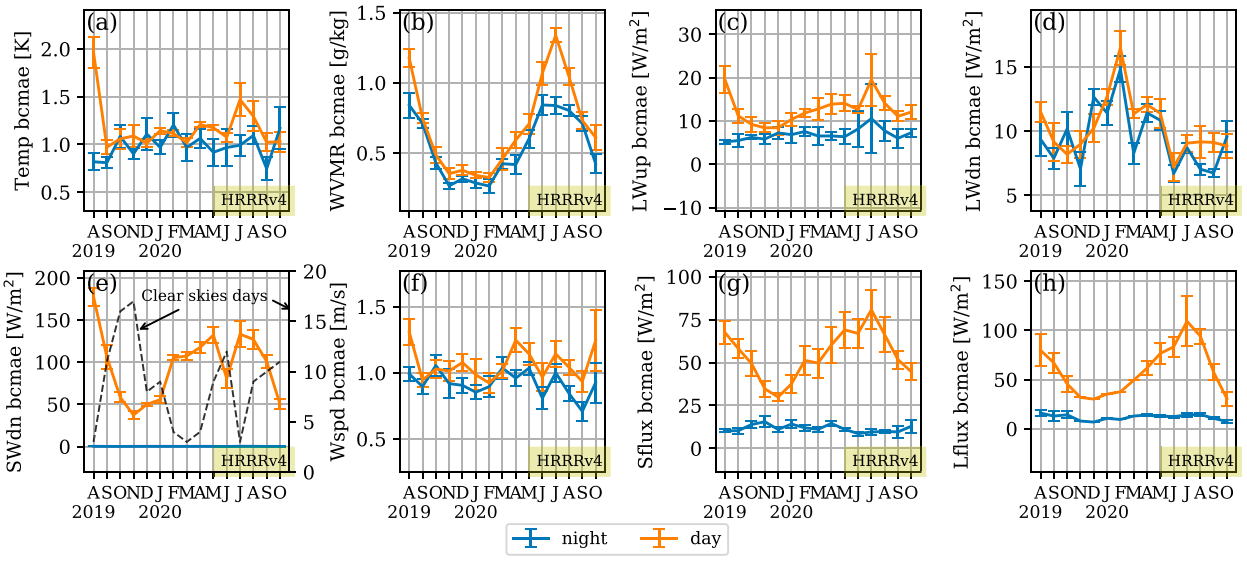

FIG. 3. As in Fig. 2, but for the BCMAE. In (e), the number of observed clear-skies days for each month is indicated by the dashed line (using the right *y* axis).

sensible heat flux was overestimated after March 2020, latent heat flux was underestimated. This is consistent with the dry daytime bias of 2-m water vapor mixing ratio for this period (Fig. 2b). The 10-m wind speed was slightly underestimated during the daytime and slightly overestimated during the nighttime (Fig. 2f).

Figure 3 shows the BCMAEs of these variables, again for HRRRv4 6-h forecasts. BCMAE values between the daytime and nighttime periods for 2-m air temperature (Fig. 3a) and downward longwave radiation (Fig. 3d) are consistent over the 15-month period. However, there are relatively large BCMAE values during the daytime of warm months for 2-m water vapor mixing ratio (Fig. 3b), upwelling longwave radiation (Fig. 3c), and 10-m wind speed (Fig. 3f). The daytime BCMAE values were markedly higher for latent (Fig. 3h) and sensible (Fig. 3g) heat fluxes; however, this is not surprising since nighttime values for those geophysical variables are generally quite small. From the observed cloud fractions, we roughly calculated the number of clear-skies days in each month shown by the dash line in Fig. 3e. We found that BCMAE of the downward shortwave radiation is highly negatively correlated with the number of clear skies in that month. If a month had more clear-skies days, the BCMAE of the downward shortwave radiation in that month was smaller. This indicates that the simulated downward shortwave radiation in HRRRv4 can be improved through improving the representation of clouds in the model.

b. *Diurnal cycle*

To study the effects of the diurnal cycles on HRRRv4's performance, we compared the simulated variables against observations at every hour over the 24-h period, using 6-h forecasts of the cold (1 December 2019–29 February 2020) and warm (1 June–31 August 2020) seasons. Similar to Figs. 2 and 3, the same metrics of each variable from all five ARM-SGP sites near to each other in Oklahoma (Fig. 1) are combined.

As expected, strong diurnal variation is evident in both the bias (Fig. 4) and BCMAE (Fig. 5) for most variables. A positive 2-m temperature bias during afternoon and evening (1200–0300 LT) is evident in the warm season but not in the cold season (Fig. 4a). The dry bias of the 2-m water vapor mixing ratio is very large (nearly 2.5 g kg^{-1}) in the daytime of the warm season (Fig. 4b); however, the 2-m water vapor is virtually bias free over the entire diurnal cycle in the cold season. The bias error of downward shortwave radiation is smaller than its BCMAE (Figs. 4e and 5e), implying that even though the downward shortwave flux is overestimated most of the time (in both the cold and warm seasons), there were times when it was underestimated. However, the bias error in the sensible heat flux is larger than its BCMAE in the warm season (Figs. 4g and 5g), suggesting that, generally speaking, the model is producing too much sensible heat flux at the surface. This is consistent with the warm/dry biases (Figs. 4a,b). Since the latent heat flux was always underestimated in the warm season (Fig. 4h), this contributes to the negative bias in the 2-m water vapor mixing ratio, and this may also lead to the underestimation of cloud and overestimation of downward shortwave radiation. The BCMAEs of sensible and latent heat fluxes are similar to the magnitude of the BCMAE in downward shortwave radiation, even though the typical magnitudes of sensible and latent heat fluxes are smaller than shortwave radiation. Last, the negative bias in the downwelling longwave flux (Fig. 4d) does not correlate well with the bias in the 2-m air temperature, suggesting that bias in air temperature is not the source of this bias in longwave radiation in both the warm and cold seasons.

c. *Forecast lead time*

Forecast skill generally decreases with the forecast lead time. The HRRR produces 48-h forecasts at four initialization

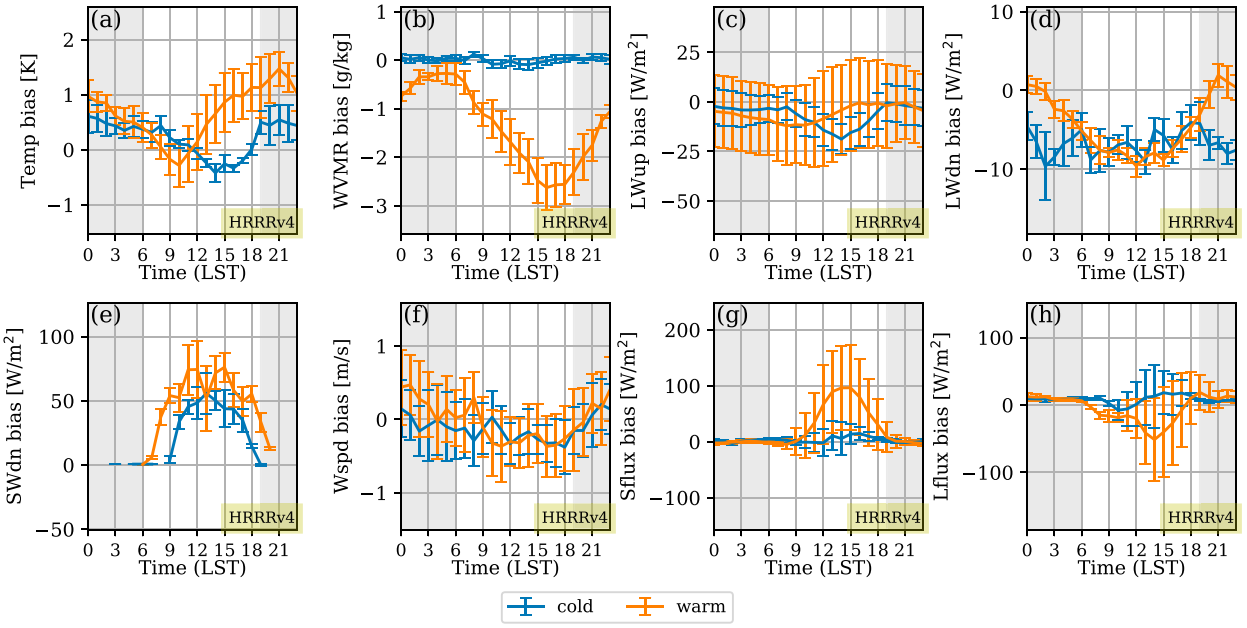

FIG. 4. Hourly bias errors of eight variables between the 6-h HRRR forecasts and observed values over the cold season (December 2019–February 2020; blue) and warm season (June–August 2020; orange). Gray shading marks the local nighttime.

times (0000, 0600, 1200, and 1800 UTC); for other initialization times, the forecasts are limited to 21 h. To avoid the mixing of the diurnal and annual cycle effects and the forecast lead time effects, metrics were calculated separately for every hour of the four-initialization time up to 48-h forecasts for the warm season (1 June–31 August 2020). The biases for the eight geophysical variables by lead time are shown in Fig. 6, and the BCMAE in Fig. 7. For example, the blue curve in Fig. 6 shows the bias errors of all forecasts initialized at 0000 UTC during the warm season as a function of forecast lead time, while all forecasts initialized at 0600, 1200, and 1800 UTC are given by the brown, green, and red curves, respectively.

As expected from the previous section, Fig. 6 shows strong variation from the diurnal cycle in bias errors for most of the eight near-surface variables, and the impact of the diurnal

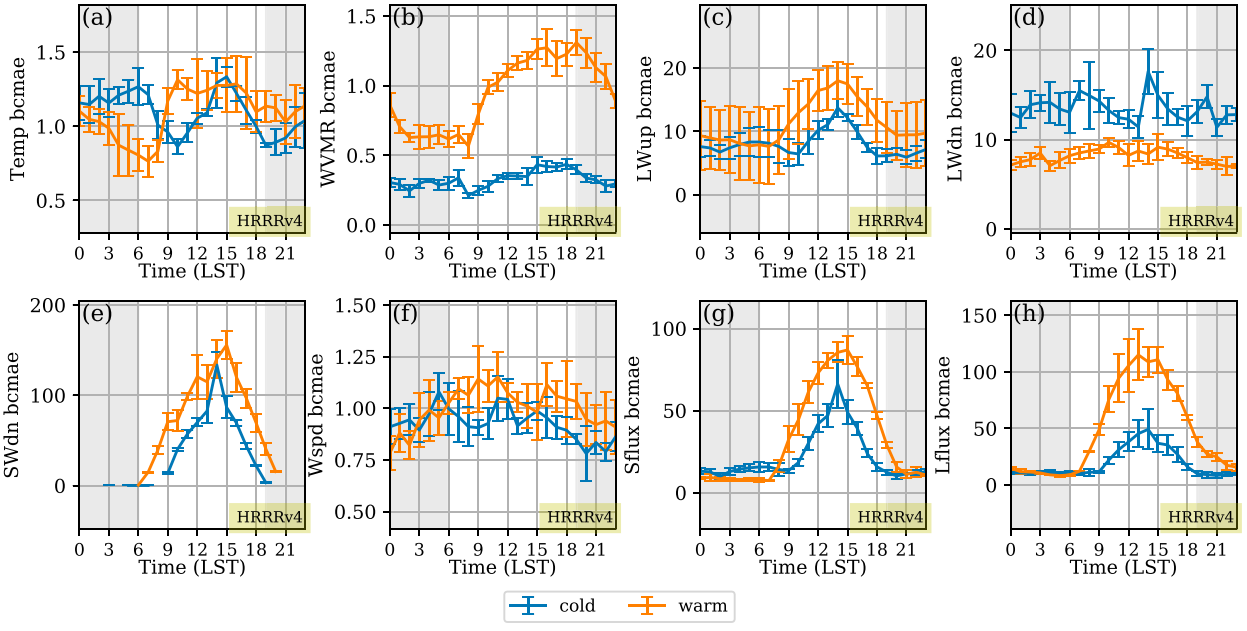

FIG. 5. As in Fig. 4, but for BCMAEs.

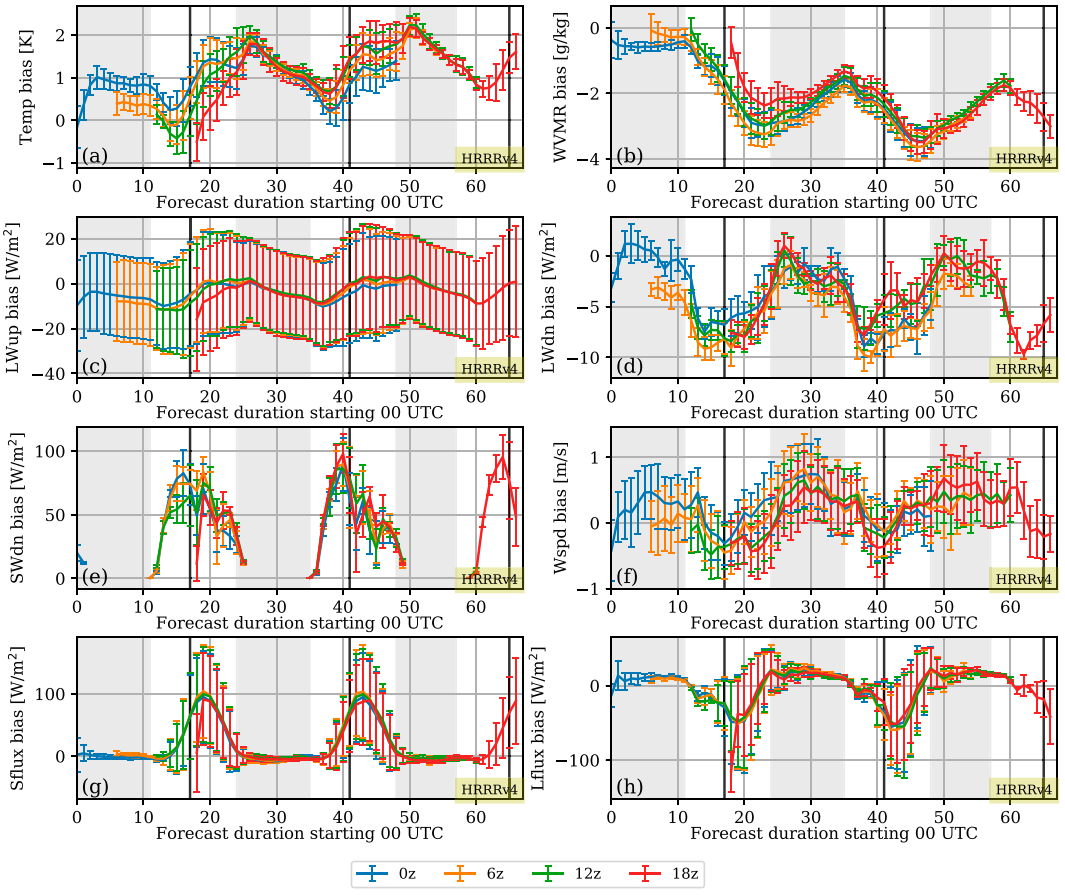

FIG. 6. Bias errors for HRRRv4 of eight variables and their variations for a warm season (1 Jun–31 Aug 2020) with forecast duration starting at 0000 UTC, for the four different initialization times of 0000, 0600, 1200, and 1800 UTC (blue, orange, green, and red, respectively). Gray areas represent nighttime periods, and the thick black vertical lines represent local noon (1200 LT).

cycle was often larger than the influence of forecast lead time. Three variables in particular show an accumulating bias error with forecast lead time in the warm season—2-m temperature (Fig. 6a), 2-m water vapor mixing ratio (Fig. 6b), and downwelling shortwave radiation (Fig. 6e), meaning that data assimilation improves the forecasts of these variables. Meanwhile, Fig. 6 shows that the bias errors of some of the variables only slightly increased with lead time (e.g., upwelling and downwelling longwave radiation fluxes, 10-m wind speed, sensible and latent heat fluxes) during the 48-h forecasts. By comparing bias errors of a same time from the different initializations, we can see that the model initialization generally improves accuracy of the simulation at short lead times (i.e., less than 3 h), especially for 2-m air temperature (Fig. 6a), 2-m water vapor mixing ratio (Fig. 6b), and downward shortwave radiation (Fig. 6e). However, the diurnal cycle effect on the bias error may still exist and is not totally eliminated through data assimilation, such as for 2-m air temperature (Fig. 6a).

However, the BCMAE gives a different perspective on skill in local variations relative to the overall bias and suggests that the model error does grow with forecast lead time for all variables (Fig. 7). This is consistent with the common assumption that forecast error increases with the increase of forecast length. Among these variables, the upwelling longwave flux (Fig. 7c), downwelling shortwave radiation (Fig. 7e), sensible heat flux (Fig. 7g), and latent heat flux Fig. 7h) showed the least sensitivity to forecast lead time, as seen by the very similar magnitude of the BCMAE at fixed times over the diurnal cycle (e.g., hours 6, 30, and 54). The increase of BCMAE in 10-m wind speed with lead time is different from the findings from Pichugina et al. (2019), which stated that the error between measured and modeled wind speed did not degrade significantly with forecast lead time. The difference may be caused by comparing wind speed errors at different levels: this study uses 10-m wind speed, while Pichugina et al. (2019) used 100-m wind speed. Also, the topographic features of these two studies are different since the SGP domain is in the relatively flat Great Plains area, while the study domain in Pichugina et al. (2019) was in more complex terrain.

d. Differences between HRRRv4 and HRRRv3

The HRRR development team gets input from a range of forecast users about shortcomings of an operational model and tries to address them via updates to the data assimilation and

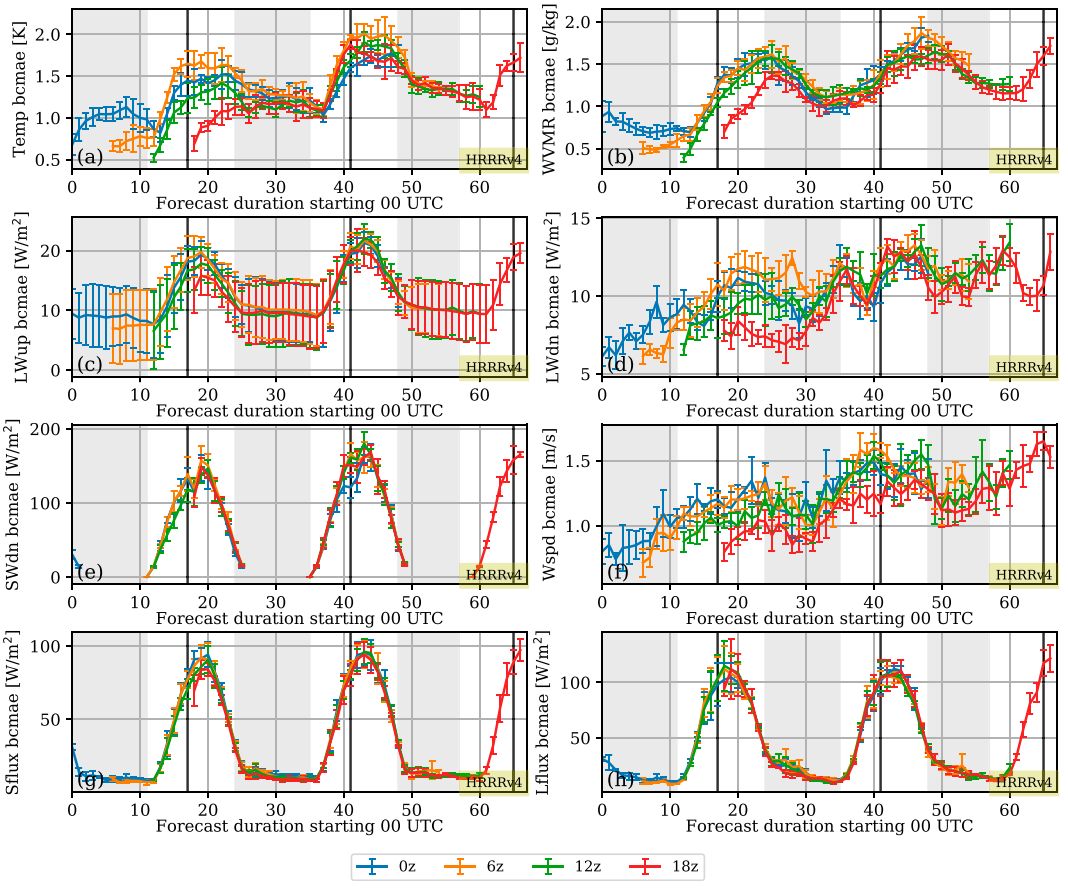

FIG. 7. As in Fig. 6, but for BCMAEs.

physics parameterizations in newer versions. The approach generally taken is outlined in Turner et al. (2020) and involves formulating hypotheses about the shortcoming and testing the impacts of designed changes (to the model parameterizations or data assimilation scheme) against observations such as the radiosondes launched by the NWS, METAR observations at the surface, radar reflectivity from the WSR-88Ds, precipitation derived from these same observations, etc. The ARM SGP observations provide a complementary way to evaluate the model differences, especially since surface turbulent fluxes are not operationally available across the nation.

The diurnal bias errors of HRRRv3 and HRRRv4 for the cold and warm seasons are presented in Fig. 8. The data from HRRR3 and HRRRv4 are "matched" (e.g., see Turner et al. 2020); if one dataset has missed a value at a time, the value of the same time from another dataset would also be removed.

The results showed that the warm bias of the 2-m air temperature in late afternoon of the cold season was corrected from HRRRv3 to HRRRv4, and the warm bias of the warm season was also corrected, especially in the morning hours. A similar daytime warm bias in the cold season for HRRRv3 was largely eliminated. This correction could come from the improvement of the SWdn from better representing subgrid clouds in HRRRv4 (see section 2b and Table 4 in D22). The high SWdn bias in the cold season was decreased dramatically (Fig. 8e); however, the bias reduction is smaller in the warm season. J22 showed a 60% reduction in excessive downward shortwave bias when averaged across seasons and using radiation measurements at a dozen sites across the United States. A lower daytime warm bias from HRRRv4 for 2-m air temperature agreed with the decrease of the high bias in the sensible heat flux (Fig. 8g). On the contrary, the dry bias errors of the 2-m water vapor mixing ratio (Fig. 8b) for the warm season are even larger in HRRRv4. J22 also noted this result, likely related to very short-range (0–1 h) low precipitation bias in HRRRv4 (see Figs. 4a, 12b, and 13b and section 2b in J22) and its data assimilation changes (an area to be addressed in the future per Fig. 3 in D22). The bias errors of downward and upward longwave radiation (Figs. 8c and 8d, respectively) were a little bit larger in HRRRv4 than in HRRRv3. The increase of the dry bias in the 2-m water vapor mixing ratio in HRRRv4 during the warm season was consistent with the increase of the low bias increment of latent heat flux (Fig. 8h). It is also possible that the increased daytime 2-m dry bias in HRRRv4 contributed to the continued deficit in cloud cover in HRRRv4.

However, the BCMAE, which provides a better measure of the model's ability to predict variability in a geophysical variable, was either close or reduced in HRRRv4 in all eight geophysical variables for most times over the diurnal cycle in both

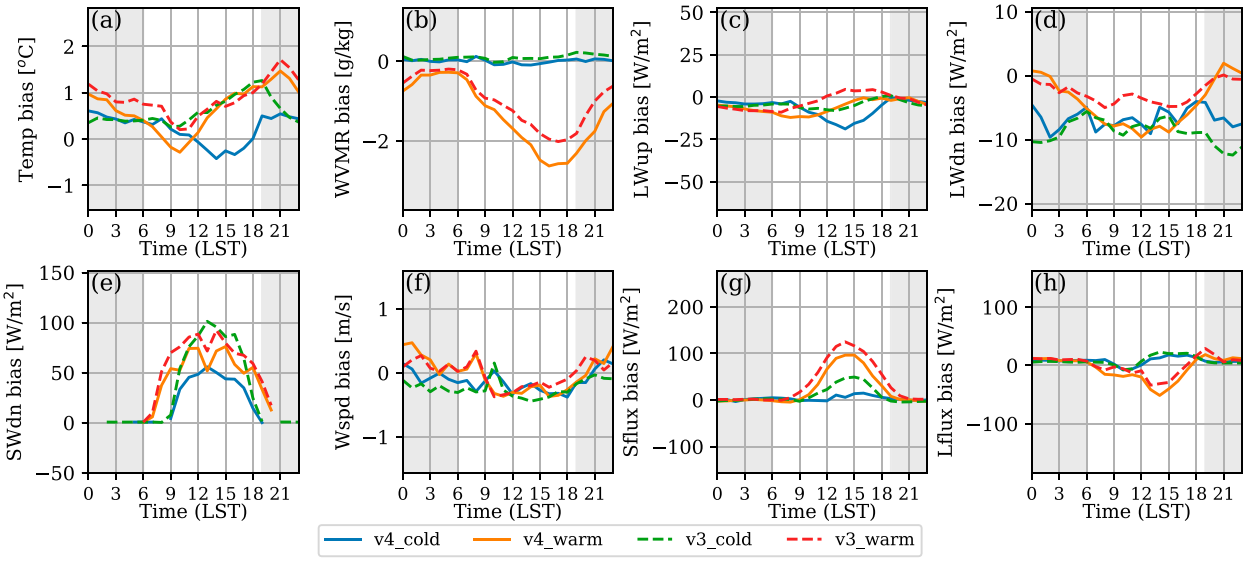

FIG. 8. Hourly bias errors of eight variables between the 6-h HRRR forecasts and observed values over cold season (December 2019–February 2020; blue for HRRRv4 and green for HRRRv3) and warm season (June–August 2020; orange for HRRRv4 and red for HRRRv3) for HRRRv3 (dashed lines) and HRRRv4 (solid lines). Gray shading marks the nighttime period.

the warm and cold seasons. The decreases in the BCMAE from HRRRv4 (relative to HRRRv3) are especially notable in the daytime 2-m temperature in both seasons (Fig. 9a), the warm-season 2-m water vapor (Fig. 9b), the warm-season upwelling longwave radiation (Fig. 9c), the warm-season downwelling shortwave radiation (Fig. 9e), and the daytime warm-season sensible and latent heat fluxes (Figs. 9g and 9h, respectively). This "across the board" improvement in the BCMAE demonstrates the superiority of HRRRv4, although the still significant magnitudes of the BCMAE and the bias in these variables indicate that more improvement to the model physics and data assimilation are certainly needed.

4. Discussion

In this section, we contrast when (by season and time of day) HRRRv4 forecasts agree closely with SGP observations and when the differences are larger to obtain insights on model design and data assimilation in HRRRv4.

First, we address the continued daytime warm bias in 2-m temperature in HRRRv4 during the warm season (Fig. 8a), slightly cooler than HRRRv3 but still positive. The main components of the surface energy budget to consider here are the downward radiative fluxes and the soil moisture driven by the ongoing cycle of 0–1-h precipitation forecasts (D22).

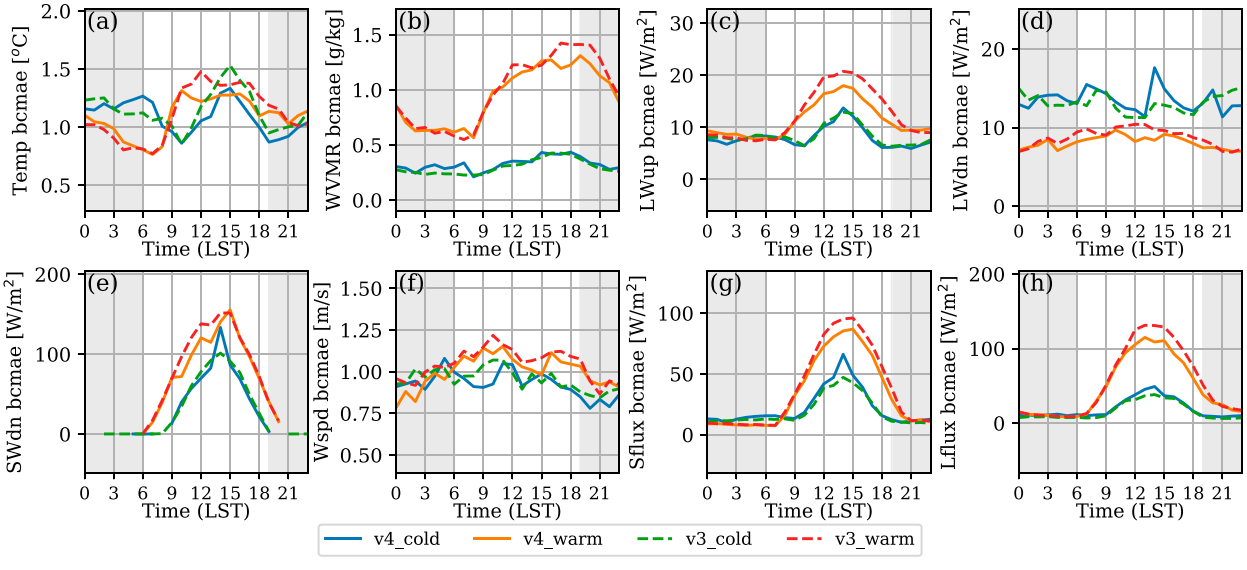

FIG. 9. As in Fig. 8, but for BCMAEs.

HRRRv4 has reduced the excessive downward shortwave bias (vs HRRRv3) over Oklahoma by about 25% (Fig. 8e) in the warm season and by about 50%–60% over the year using 14 radiation stations across the United States (Fig. 9 in J22). At the same time, a 2-m moisture dry bias in the warm season, especially during daytime, became more pronounced with HRRRv4 (Fig. 8b) (Fig. 12 in J22). We attribute this to reduced 0–1-h precipitation in HRRRv4 from data assimilation changes, resulting in a subsaturation bias that led to a dry soil moisture bias (see Fig. 4a and section 2b in J22, and the end of section 2d in D22). This precipitation deficit in the 0–1-h forecast is critical for the evolution of the HRRR soil moisture (Benjamin et al. 2022a). This attribution is also consistent with the lower bias in the latent heat flux (Fig. 8h), with the lower soil moisture biases due to less precipitation. It is possible that this daytime dry bias contributed to still-deficient daytime clouds in HRRRv4. As shown in section 3, the agreement with SGP observations generally improved with HRRRv4 compared with HRRRv3, so our remaining discussion below is in the context of improved downward solar radiation with HRRRv4 but with drier soil conditions due to the change in the data assimilation implemented in HRRRv4.

Air temperature, water vapor, and cloud properties (i.e., height, optical depth, and thermodynamic phase) all affect the simulation of the downward longwave radiation. A high bias of air temperature can result in an increase in the downward longwave radiation. In HRRRv4, the high bias trend of 2-m air temperature in the warm season is similar to the low bias trend of downward longwave radiation to some extent. This means that a high bias of 2-m air temperature did lead to a less low bias of downward longwave radiation. A low bias in 2-m water vapor mixing ratio could result in fewer clouds and thus, less downward longwave radiation, and this is consistent with our results. This means that the dry bias of 2-m water vapor mixing ratio dominated the low bias of downward longwave radiation. However, the diurnal low bias of downward longwave radiation in the warm season (Fig. 8d) is not in the same trend as the diurnal bias in 2-m water vapor mixing ratio. We hypothesize that the model is underestimating cloud longwave radiative flux (e.g., either with a low bias in cloud fraction or optical depth, or by placing the cloud too high), which results in an underestimation of downward longwave radiation. The first two (i.e., low bias in cloud fraction or cloud optical depth) are consistent with the overestimation of downward shortwave radiation. By comparing observations from *GOES-13*, Griffin et al. (2017) also showed that the HRRRv3 forecasts contained fewer cloud objects than were observed. From comparisons of HRRRv3 and HRRRv4 in this study (e.g., Fig. 8e), the cloud fraction in HRRRv4 was improved but still is underestimated.

Downward shortwave radiation is the geophysical variable still overestimated in HRRRv4 despite significant improvement from HRRRv3. Clouds are the main factor that affects the simulation of solar radiation, and too few clouds (e.g., an underestimate of cloud fraction or cloud optical depth) leads to a positive bias in the downward solar radiation. Since the model overestimates shortwave radiation, we conclude that its cloud fraction was also underestimated. This is consistent with the underestimation of downward longwave radiation. Another fact supporting the clouds-related issue is that BCMAE for downward shortwave radiation was much smaller in months with clearer-skies days than months with fewer clear-sky days (Fig. 3e). It is still possible that the deficit of clouds is forced by the daytime warm-season 2-m warm and dry bias, which may, in turn, be a consequence of a dry soil bias from the 0–1-h precipitation deficiency related to the HRRRv4 data assimilation.

The monthly and hourly bias errors of 10-m wind speed showed that the model underestimated winds by about 0.2 m s^{-1} during daytime throughout the year, although there is some spatial variability in the results as the 10-m wind speed bias errors of the five sites are scattered (not shown). This slight underestimation of daytime 10-m wind speed is consistent with the surface verification of Fovell and Gallagher (2020) and Pichugina et al. (2019).

Sensible and latent heat fluxes are a challenge to simulate, as they are impacted by many factors both within the land surface and the atmosphere. The large daytime warm-season flux biases are consistent with the 2-m warm and dry bias. We also note the challenge in comparing the HRRR against surface flux observations from representativeness of flux observations, which have relatively small footprints (of order 10 s of m) compared with the 3-km grid cell, as mentioned earlier in this paper. Many studies have shown that measuring surface fluxes can be difficult (e.g., LeMone et al. 2019), and their spatial heterogeneities are also strong (e.g., Tang et al. 2019).

In our study, we found some consistency with Lee et al. (2019), but there are different findings as well. From the cold-season diurnal cycle statistics, it was found that the HRRRv4 slightly underestimated 2-m air temperature during the daytime and overestimated it during the nighttime. However, Lee et al. (2019) reported the opposite finding for their cold-season study using HRRRv2, which was conducted in northern Alabama. Both studies showed that the surface fluxes from the HRRR had large differences with the observations, although here, we show lower errors in HRRRv4 (Figs. 8g,h and 9g,h). Also, in HRRRv4, the EDMF approach and subgrid cloud parameterization are included to improve the representation of nonlocal mixing and shortwave radiation.

By investigating the model error as a function of forecast lead time, this study found that the bias errors in the forecast of the near-surface variables largely depend on the diurnal cycle of the surface variables themselves, which is the same as findings from Zhang et al. (2013). Yue and Gebremichael (2020) also showed that the effect of lead times was not obvious as there was no systematic difference in accuracy among the 2–18-h lead-time precipitation forecasts. However, the BCMAE statistics (Fig. 7) demonstrate that the forecast errors in all variables do have an upward trend with forecast lead time, although some variables (e.g., upwelling longwave, latent flux, and sensible flux) have relatively small changes in BCMAE over the forecast time.

5. Conclusions

The simulated HRRRv4 near-surface geophysical variables, including wind, humidity, temperature, surface latent and sensible fluxes, and longwave and shortwave radiative fluxes, were compared against ARM SGP observations in this study

using statistical metrics including bias and bias-corrected mean absolute error (BCMAE). From these systematic comparisons across different time scales and leading times, we have some conclusions about the performance of HRRRv4 over the SGP domain in Oklahoma.

The differences between HRRRv4 forecasts and SGP observations in Oklahoma were small for many variables and seasons, but for surface heat fluxes were sometimes larger in warm seasons (e.g., cold bias in fall and dry bias in summer) and in the daytime when sensible and latent heat fluxes had larger values. The monthly results of hourly comparisons showed that the model's performance was relatively stable over the 15 months, and daytime and nighttime biases were different. However, some of the variables showed seasonal variations, such as the 2-m water vapor mixing ratio. The results also showed that nighttime 2-m air temperature was overestimated over the 15 months, and daytime 2-m air temperature was slightly underestimated in the fall months of the 15 months.

The hourly results of the warm and cold seasons further demonstrated that diurnal cycle has, as expected, a large impact on the bias of the simulated near-surface variables. The 2-m air temperature has a warm bias during late afternoon in the warm season and a slight cold bias during late afternoon in the cold season. The low bias of downward longwave radiation was larger during the daytime than the nighttime in the warm season, but the low bias was relatively constant during the cold season.

The monthly and hourly results showed that HRRRv4 has a 2-m dry bias in the warm season, and additional improvements in simulating the surface heat fluxes are required to match the in situ observations. Also, the model's performance in simulating the radiative and surface heat fluxes implies that the representation of clouds needs to be improved in the model. Since this study only investigated some of the near-surface variables, we cannot specifically identify which cloud properties have deviations.

The impact of the forecast lead time on the model's performance shows the diurnal cycle is the dominant factor in controlling forecast error, although the bias-corrected mean absolute error showed the upward trend over the increase in lead time. The results also demonstrated that data assimilation does improve HRRR's performance on most of the variables, but not all of the variables.

Comparisons were also made between forecasts from HRRRv3 and HRRRv4 and the SGP observations, showing a general improvement from HRRRv4 for these boundary layer variables. Almost all variables were improved in the cold season, and the same was true for the warm season except for 2-m water vapor. Reasons for these improvements were described in the paper, including improvements in HRRRv4 model physics but also with reduced short-range precipitation in HRRRv4 from data assimilation changes described by J22. Overall, the HRRRv4 compared with HRRRv3 performance shown here was largely consistent with the differences with national-scale verification also shown by J22. The results of this paper will guide new developments in model physics and data assimilation for NOAA convection-allowing scale modeling.

Acknowledgments. The authors thank Dr. Ryann Wakefield of NOAA Global Systems Laboratory for an insightful review of this paper. Support for this analysis was provided via Grant 89243019SSC000034 provided by the DOE Atmospheric System Research (ASR) program and by the NOAA Atmospheric Science for Renewable Energy (ASRE) program. Tatiana G. Smirnova and Siwei He are supported by NOAA Cooperative Agreement NA22OAR4320151.

Data availability statement. The in situ observed data of the five sites around the SGP domain are downloaded from the ARM web page (https://www.arm.gov/). The HRRRv3 and HRRRv4 data are available through Amazon Web Services (https://registry.opendata.aws/noaa-hrrr-pds/).

REFERENCES

Angevine, W. M., J. Olson, J. J. Gristey, I. Glenn, G. Feingold, and D. D. Turner, 2020: Scale awareness, resolved circulations, and practical limits in the MYNN–EDMF boundary layer and shallow cumulus scheme. *Mon. Wea. Rev.*, **148**, 4629–4639, https://doi.org/10.1175/MWR-D-20-0066.1.

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.

——, 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.

——, T. G. Smirnova, E. P. James, L.-F. Lin, M. Hu, D. D. Turner, and S. He, 2022a: Land–snow assimilation including a moderately coupled initialization method applied to NWP. *J. Hydrometeor.*, **23**, 825–845, https://doi.org/10.1175/JHM-D-21-0198.1.

——, and Coauthors, 2022b: Inland lake temperature initialization via coupled cycling with atmospheric data assimilation. *Geosci. Model Dev.*, **15**, 6659–6676, https://doi.org/10.5194/gmd-15-6659-2022.

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

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

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

——, and ——, 2022: An evaluation of surface wind and gust forecasts from the High-Resolution Rapid Refresh. *Wea. Forecasting*, **37**, 1045–1068, https://doi.org/10.1175/WAF-D-21-0176.1.

Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, https://doi.org/10.1175/WAF-D-17-0144.1.

Griffin, S. M., J. A. Otkin, C. M. Rozoff, J. M. Sieglaff, L. M. Cronce, C. R. Alexander, T. L. Jensen, and J. K. Wolff, 2017: 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.

He, S., T. G. Smirnova, and S. G. Benjamin, 2021: Single-column validation of a snow subgrid parameterization in the Rapid Update Cycle Land-Surface Model (RUC LSM). *Water Resour. Res.*, **57**, e2021WR029955, https://doi.org/10.1029/2021WR029955.

Homer, C. G., and Coauthors, 2015: Completion of the 2011 National Land Cover Database for the conterminous United States—Representing a decade of land cover change information. *Photogramm. Eng. Remote Sens.*, **81**, 345–354.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, https://doi.org/10.1029/2008JD009944.

Ikeda, K., M. Steiner, 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.

James, E. P., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part II: Forecast performance. *Wea. Forecasting*, **37**, 1397–1417, https://doi.org/10.1175/WAF-D-21-0130.1.

Koster, R. D., and Coauthors, 2004: Regions of strong coupling between soil moisture and precipitation. *Science*, **305**, 1138–1140, https://doi.org/10.1126/science.1100217.

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.

LeMone, M. A., and Coauthors, 2019: 100 years of progress in boundary layer meteorology. *A Century of Progress in Atmospheric and Related Sciences: Celebrating the American Meteorological Society Centennial, Meteor. Monogr.*, No. 59, Amer. Meteor. Soc., https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0013.1.

Morcrette, C. J., and Coauthors, 2018: Introduction to CAUSES: Description of weather and climate models and their near-surface temperature errors in 5 day hindcasts near the southern Great Plains. *J. Geophys. Res. Atmos.*, **123**, 2655–2683, https://doi.org/10.1002/2017JD027199.

Olson, J. B., J. S. Kenyon, W. A. Angevine, J. M. Brown, M. Pagowski, and K. Sušelj, 2019a: A description of the MYNN-EDMF scheme and the coupling to other components in WRF-ARW. NOAA Tech Memo. OAR GSD 61, 42 pp., https://doi.org/10.25923/n9wm-be49.

——, and Coauthors, 2019b: Improving wind energy forecasting through numerical weather prediction model development. *Bull. Amer. Meteor. Soc.*, **100**, 2201–2220, https://doi.org/10.1175/BAMS-D-18-0040.1.

——, T. Smirnova, J. S. Kenyon, D. D. Turner, J. M. Brown, W. Zheng, and B. W. Green, 2021: A description of the MYNN surface-layer scheme. NOAA Tech. Memo. OAR GSL-67, 26 pp., https://doi.org/https://doi.org/10.25923/f6a8-bc75.

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 Basin. *J. Appl. Meteor. Climatol.*, **58**, 1633–1656, https://doi.org/10.1175/JAMC-D-18-0244.1.

Sisterson, D. L., R. A. Peppler, T. S. Cress, P. J. Lamb, and D. D. Turner, 2016: The ARM Southern Great Plains (SGP) site. *The Atmospheric Radiation Measurement Program: The First 20 Years*, *Meteor. Monogr.*, No. 57, Amer. Meteor. Soc., https://doi.org/10.1175/AMSMONOGRAPHS-D-16-0004.1.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and D. Kim, 2000: Parameterization of cold-season processes in the MAPS land-surface scheme. *J. Geophys. Res.*, **105**, 4077–4086, https://doi.org/10.1029/1999JD901047.

——, ——, ——, and J. S. Kenyon, 2016: Modifications to the Rapid Update Cycle Land Surface Model (RUC LSM) available in the Weather Research and Forecasting (WRF) Model. *Mon. Wea. Rev.*, **144**, 1851–1865, https://doi.org/10.1175/MWR-D-15-0198.1.

Tang, S., S. Xie, M. Zhang, Q. Tang, Y. Zhang, S. A. Klein, D. R. Cook, and R. C. Sullivan, 2019: Differences in eddy-correlation and energy-balance surface turbulent heat flux measurements and their impacts on the large-scale forcing fields at the ARM SGP site. *J. Geophys. Res. Atmos.*, **124**, 3301–3318, https://doi.org/10.1029/2018JD029689.

Turner, D. D., and R. G. Ellingson, Eds., 2016: *The Atmospheric Radiation Measurement (ARM) Program: The First 20 Years. Meteor. Monogr.*, No. 57, Amer. Meteor. Soc., 504 pp.

——, and W. G. Blumberg, 2019: Improvements to the AERIoe thermodynamic profile retrieval algorithm. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **12**, 1339–1354, https://doi.org/10.1109/JSTARS.2018.2874968.

——, and Coauthors, 2004: The QME AERI LBLRTM: A closure experiment for downwelling high spectral resolution infrared radiance. *J. Atmos. Sci.*, **61**, 2657–2675, https://doi.org/10.1175/JAS3300.1.

——, 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.

Wagner, T. J., P. M. Klein, and D. D. Turner, 2019: A new generation of ground-based mobile platforms for active and passive profiling of the boundary layer. *Bull. Amer. Meteor. Soc.*, **100**, 137–153, https://doi.org/10.1175/BAMS-D-17-0165.1.

Yang, F., H.-L. Pan, S. K. Krueger, S. Moorthi, and S. J. Lord, 2006: Evaluation of the NCEP Global Forecast System at the ARM SGP site. *Mon. Wea. Rev.*, **134**, 3668–3690, https://doi.org/10.1175/MWR3264.1.

Yue, H., and M. Gebremichael, 2020: Evaluation of High-Resolution Rapid Refresh (HRRR) forecasts for extreme precipitation. *Environ. Res. Commun.*, **2**, 065004, https://doi.org/10.1088/2515-7620/ab9002.

Zhang, H., Z. Pu, and X. Zhang, 2013: Examination of errors in near-surface temperature and wind from WRF numerical simulations in regions of complex terrain. *Wea. Forecasting*, **28**, 893–914, https://doi.org/10.1175/WAF-D-12-00109.1.