On the Importance of Regime-Specific Evaluations for Numerical Weather Prediction Models as Demonstrated Using the High-Resolution Rapid Refresh (HRRR) Model

TEMPLE R. LEE, SANDIP PAL, RONALD D. LEEPER, TIM WILSON, HOWARD J. DIAMOND, TILDEN P. MEYERS, AND DAVID D. TURNER

ABSTRACT: The scientific literature has many studies evaluating numerical weather prediction (NWP) models. However, many of those studies averaged across a myriad of different atmospheric conditions and surface forcings that can obfuscate the atmospheric conditions when NWP models perform well versus when they perform inadequately. To help isolate these different weather conditions, we used observations from the U.S. Climate Reference Network (USCRN) obtained between 1 January and 31 December 2021 to distinguish among different near-surface atmospheric conditions and surface forcings that can obfuscate the performance of these biases helps the meteorological community understand, modify, and improve the physical parameterization schemes used therein, which we anticipate will lead to more accurate weather forecasts. Oftentimes when model evaluations are conducted many different atmospheric conditions, ranging from highly convective to very stable conditions, are averaged together. For example, the High-Resolution Rapid Refresh (HRRR), which is a 3-km convection-allowing model that for nearly 10 years has been used to support operational weather forecasting needs in the United States (e.g., Benjamin et al. 2016; Dowell et al. 2022; James et al. 2022), has previously been evaluated across multiple months and seasons. These studies identified biases between the model and observations and indicated several model-data mismatches (MDMs) in, for example the HRRR’s near- and subsurface meteorological fields (e.g., Min et al. 2021; Patel et al. 2021; Lee et al. 2023, hereafter L23), surface fluxes (e.g., Lee et al. 2019; He et al. 2023), precipitation (e.g., Yue and Gebremichael 2020; Duda and Turner 2021; English et al. 2021), and atmospheric thermodynamic and kinematic profiles (e.g., Fovell and Gallagher 2020; MacDonald and

1. Introduction

Evaluating operational numerical weather prediction (NWP) models is critical for identifying weather conditions in which the models perform well in addition to the scenarios where the models perform inadequately. Knowledge of the conditional nature of these biases helps the meteorological community understand, modify, and improve the physical parameterization schemes used therein, which we anticipate will lead to more accurate weather forecasts. Oftentimes when model evaluations

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

Corresponding author: Temple R. Lee, temple.lee@noaa.gov

DOI: 10.1175/WAF-D-23-0177.1

© 2024 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
Nowotarski 2023). Whereas these studies provide insights into model performance on seasonal and subseasonal time scales (e.g., Lee et al. 2019; Min et al. 2021), the key findings from these studies were reported irrespective of weather conditions.

Previous studies have also yet to address how model performance varies due to both gradual and rapid changes in near-surface atmospheric conditions caused by different land surface forcings (e.g., flash droughts, flash floods), airmass changes (e.g., during frontal passages and dryline passages), etc. We argue that model performance strongly depends on the atmospheric conditions under investigation, and model error diagnostics should be performed accordingly by classifying weather conditions without performing bulk statistics on model errors. Positive biases under one set of ambient atmospheric conditions and negative biases under another set of conditions, when averaged together, yield small mean biases. Consequently, these “global means” are not fully indicative of model performance and can obfuscate our ability to isolate the specific conditions in which the model biases are largest versus when model biases are smallest. In the example of the HRRR, which is the focus of the present study, knowledge of regime-specific model performance (i.e., MDMS as a function of different weather conditions) is anticipated to provide targeted areas for additional research and model development and help guide improvements to the land surface and atmospheric boundary layer (ABL) parameterizations in subsequent HRRR iterations, i.e., the Rapid Refresh Forecast System (RRFS; Dowell et al. 2022).

In this study, we use high-quality observations obtained from the U.S. Climate Reference Network (USCRN) over a 1-yr period (i.e., 1 January 2021–31 December 2021) to evaluate forecasts of air temperature, incoming shortwave radiation, and soil moisture from the HRRR over three subsets of different conditions (i.e., different near-surface heating rates, radiative regimes, and soil moisture regimes) to perform a weather-specific model evaluation, as summarized in Fig. 1. We then compare the results with model errors resulting from averaging across the entire range of atmospheric conditions observed during the 1-yr study period, which was previously reported in numerous previous studies, for example, L23 and which is the conventional model evaluation approach used to examine performance of operational NWP models.

2. Datasets

a. USCRN

The USCRN currently comprises 114 stations installed at carefully selected sites in the contiguous United States (i.e., CONUS) and has been operating continuously across this domain since 2008 (Fig. 1a). The stations are spaced at the spatial resolution necessary to capture the temporal variability in temperature and precipitation trends over CONUS (Diamond et al. 2013). The spatial resolution of the USCRN stations over CONUS was determined using observations from thousands of stations in the Cooperative Observing Network (Vose and Menne 2004; Vose 2005; Diamond et al. 2013). At each USCRN station are in situ measurements of air temperature ($T$) at 1.5 m above ground level (AGL), surface skin temperature, precipitation, incoming shortwave radiation ($SW_d$), and soil temperature and soil moisture (SM) at 5, 10, 20, 50, and 100 cm below the surface. Soil temperature and soil moisture are available at all five of these depths at 90 of the USCRN stations, whereas 5- and 10-cm measurements are available from 23 of the USCRN stations. Due to solid rock at the station at Torrey, Utah, there are no soil measurements available from that site (e.g., Bell et al. 2013).

The measurements from USCRN are of high quality; for example, $T$ has an accuracy of ±0.3°C over the range from −50° to +50°C, and the raw $T$ data have a resolution of 0.01°C (NOAA/NESDIS 2007). As this network has been well-described in previous work, we refer the reader to previous studies (e.g., Bell et al. 2013; Diamond et al. 2013; L23) that provide additional information about USCRN, including details about site selection, instrument type, data quality control and quality assurance procedures, etc. In the analyses in the present study, we focus on USCRN’s observations of $T$, $SW_d$, and 5-cm SM.

b. HRRR

We used model output, available in GRIB2 format, from version 4 of the HRRR (hereinafter referred to as the HRRRv4) which has a 3-km grid spacing and 51 vertical layers that extend up to 15 hPa (e.g., Dowell et al. 2022). HRRRv4 uses the Rapid Update Cycle (RUC) land surface model (LSM) (e.g., Smirnova et al. 2016), Mellor–Yamada–Nakanishi–Niino eddy diffusivity mass flux (MYNN-EDMF) boundary layer scheme (Nakanishi and Niino 2004, 2009; Olson et al. 2019), MYNN surface-layer scheme (Olson et al. 2021), Thompson microphysics scheme (e.g., Thompson and Eidhammer 2014), and the Rapid Radiative Transfer Model Global for radiation (e.g., Iacono et al. 2008). We note, though, that the MYNN surface-layer scheme outputs 2-m $T$ whereas USCRN measures $T$ at 1.5 m AGL, and we do not explicitly account for these height discrepancies in our analyses. More details about the surface-layer scheme and other parameterization schemes used in the HRRR have been reported in, for example Benjamin et al. (2016), Dowell et al. (2022), and James et al. (2022).

3. Methods

The RUC LSM used in the HRRRv4 has 21 land-cover categories. Whereas 12 of these categories are represented at the USCRN stations, most of the USCRN stations have either the cropland or grassland land-cover type as was noted in L23. Eight stations have the water land-cover classification due to their proximity to lakes or oceans (Fig. 1a). Because the grid cells with the water land-cover classification have different characteristics (e.g., saturated soils) in the RUC LSM than the other land-cover classifications, we did not use the eight USCRN stations with the water land-cover classification in any analyses.

In this study we focus on the performance of the HRRRv4’s 18-h forecast to illustrate the different biases resulting from different near-surface atmospheric conditions. The 18-h forecast is run every hour; consequently, there are 8760 1-h time periods considered during the 1-yr study period at each of the
USCRN stations. This forecast cycle is less sensitive to the model’s initial conditions and data assimilation than earlier forecast hours; therefore, evaluating the 18-h forecast is arguably a better assessment of the model physics than earlier forecast hours. However, using the same evaluations for earlier forecast hours (i.e., the 1-h forecast from the HRRRv4) yields similar results, indicating that our results are not strongly sensitive to our choice of forecast hour. We revisit this point in section 4.

We obtained HRRRv4 output from the NOAA Open Data Dissemination (NODD) program via the Amazon Web Services (AWS), and we used the procedure discussed in Blaylock et al. (2017). Following L23, we omitted observations of a given variable from a given station if >25% of the total expected observations within the experimental time frame were missing. For example, if Station X had 70% data completion over the period of record for 5-cm SM, but Station X had 80% data completion for T, we would omit the 5-cm SM

---

**Fig. 1.** (a) USCRN stations in CONUS (white squares). Stations that are omitted because of the water land-cover classification have an orange circle (cf. section 3). (b) Percentiles of $dT/dt$, $SW_d$, and $SM_{05}$ calculated across all USCRN stations for 2021 are shown. To show all variables on the same graph, $dT/dt$, $SW_d$, and $SM_{05}$ were multiplied by 10, 0.1, and 100, respectively. The black circles represent the 50th percentile; the error bars extend outward to the 25th and 75th percentiles; and the red circles are the 10th and 90th percentiles. (c) We summarize our model evaluation approach. Source of the background map in (a): GoogleEarth.
observations from Station X from all further analyses, but we would still use the \( T \) observations from Station X. Incomplete data records are most common for 5-cm SM. In 2021, 70\% of the USCRN stations have \( 75\% \) data completion for 5-cm SM, whereas 99\% of the USCRN stations have \( >75\% \) data completion for \( T \), and 98\% of the USCRN stations have \( >75\% \) data completion for \( SW_d \). The smaller SM data completion occurs because of soils freezing during the winter months at northern latitudes.

We then used the USCRN datasets to differentiate among meteorological regimes. In the present study we distinguish among ABL regimes using an encroachment model (e.g., Carson 1973; Tennekes 1973; Batchvarova and Gryning 1991). In encroachment models, the dynamics of turbulent entrainment are neglected, and the ABL growth rate is approximated as a function of the daytime heating rate \( (dT/dt) \) and ABL lapse rate. Ever since encroachment models were developed more than 50 years ago, they have successfully been used for example in recent studies on ABL growth and evolution (e.g., Pal et al. 2010, 2013). We computed \( dT/dt \) using the USCRN \( T \) observations to distinguish among days with large heating rates, and thus typically deeper ABLs, from days with small daytime heating rates and thus typically more shallow ABLs. We computed \( dT/dt \) as the \( T \) difference between 1400 and 0800 LST. Over the 1-yr study period, there was a large range in the observed daily values of \( dT/dt \), as the daily \( dT/dt \) across all 114 USCRN stations in 2021 ranged from \(-2.5\) to \(4.8\) K h\(^{-1}\). Cases in which \( T \) decreased during the daytime (i.e., days with \( dT/dt < 0 \) K h\(^{-1}\)) occurred on about 5\% of all days and oftentimes had cold front passages that resulted in daytime \( T \) decreases. We performed sensitivity tests using different start and end times postsunrise and presunset, respectively, and found that our results remained independent of our choice for the starting hour or ending hour for computing \( dT/dt \) (not shown).

Additionally, we classified \( dT/dt \) into four different regimes based on histogram analyses of the percentiles, which we computed across all USCRN stations, to capture a range of magnitudes of near-surface forcings. We then determined the MDM by calculating the mean bias error (MBE) as the difference between the modeled and observed values for four \( dT/dt \) classifications:

1) \(<25\text{th percentile of } dT/dt \>
2) \(25\text{th–50th percentile of } dT/dt \>
3) \(50\text{th–75th percentile of } dT/dt \>
4) \(>75\text{th percentile of } dT/dt \>

As shown in Fig. 1b, the 25th and 75th percentiles, computed across stations, for \( dT/dt \) were 0.57 and 1.40 K h\(^{-1}\), respectively. The largest \( dT/dt \) occurred over the semiarid U.S. Southwest and southern Great Plains where the 25th \( dT/dt \) percentile was \(-1\) K h\(^{-1}\), and the 75th \( dT/dt \) percentile was \(>2\) K h\(^{-1}\) (Figs. 2a–c). The smallest \( dT/dt \) generally occurred at coastal sites in the eastern United States. Once we
computed \( \frac{dT}{dt} \), we then analyzed the MBEs for the different \( \frac{dT}{dt} \) as a function of local standard time at each station. To facilitate a comparison with L23 (i.e., their Fig. 8), we report the results as a function of UTC rather than in local time. When considering the nighttime, i.e., roughly from 0200 to 1200 UTC, we used the \( \frac{dT}{dt} \) percentile for the upcoming daytime period. For the investigation of the spatial variability of the MBEs, we computed the MBE as a function of the difference between the HRRR and USCRN observations obtained between 1800 and 0000 UTC to represent the late afternoon.

FIG. 3. Mean diurnal MBE cycle of (a) \( T \), (b) \( SW_d \), and (c) \( SM_{05} \) for the different percentiles of \( \frac{dT}{dt} \) observed across all USCRN stations. (d)(f)(g)(i) As in (a)(c), but for the different \( SW_d \) and \( SM_{05} \) percentiles, respectively. The purple, blue, orange, and red lines represent <25th, 25th–50th, 50th–75th, and >75th percentiles, respectively; the thick black line represents the mean across the entire study period.
mean values. Furthermore, we computed the means of the MBEs over the entire year to assess their spatial variability. We repeated the analyses discussed above for different magnitudes of \( \frac{dT}{dt} \), but used USCRN’s observations of mean daily shortwave radiation (\( SW_d \)) and mean daily 5-cm soil moisture (\( SM_{05} \)) to quantify the MDM across different radiative regimes and SM states, respectively, and compared these findings against the MDM over the entire study period. Over the study period, the 25th and 75th percentiles for \( SW_d \) (\( SM_{05} \)) were 102 W m\(^{-2}\) (0.11 m\(^3\) m\(^{-2}\)) and 250 W m\(^{-2}\) (0.31 m\(^3\) m\(^{-2}\)), respectively (Figs. 1b,c). The \( SW_d \) and \( SM_{05} \) percentiles exhibited expected north–south and west–east gradients, respectively, with the largest \( SW_d \) and smallest \( SM_{05} \) occurring over the U.S. Southwest (Figs. 2d–i).

4. Results and discussion

a. Diurnal biases

1) Air temperature

When distinguishing among different heating rates, we find substantial differences in the mean diurnal cycle of \( T \) MBE (Fig. 3a). For low heating rates (i.e., \( dT/dt < 25 \text{th percentile} \)), the \( T \) MBE is very small (\( \sim 0^\circ \text{C} \)) during the nighttime but \( 2^\circ \text{C} \) during the daytime which is largely in contrast to the subsets of days with larger heating rates as well as the mean across all cases (i.e., thick black line in Fig. 3). Days with the largest \( dT/dt \) (i.e., \( dT/dt > 75 \text{th percentile} \)) have the largest \( T \) MBE during the nighttime. On this subset of days, the MBE peaks \( \sim 4^\circ \text{C} \) at 1200 UTC but decreases during the daytime to \( \sim 1^\circ \text{C} \). These values are much different than the mean diurnal cycle in MBE computed across all days, whereby the HRRRv4 has a positive (i.e., warm) bias of \( \sim 1^\circ \text{C} \) during the nighttime and slightly negative (i.e., cool) bias between 1500 and 1800 UTC. Repeating these analyses using HRRR forecasts closer to the model’s initialization, i.e., using the 1-h HRRR forecast rather than the 18-h HRRR forecast, shows than the MBE amplitude is smaller in the 1-h HRRR forecast than in 18-h HRRR forecast, which is consistent with findings from L23 (their Fig. 8). However, the MBE for the different \( dT/dt \) percentiles follows the same diurnal pattern in the 1-h forecast as the 18-h forecast, suggesting that our results are not strongly sensitive to the choice of HRRR forecast hour that is evaluated.

Additionally, we find that varying \( dT/dt \) has substantial impacts on \( SW_d \) MBE. On days with the smallest \( dT/dt \), the MBE reaches a daytime maximum of about 170 W m\(^{-2}\) (Fig. 3b). Conversely, days with the largest \( dT/dt \) yielded maximum MBEs of 100 W m\(^{-2}\). In contrast with these findings and, perhaps not surprisingly, the impacts of different \( dT/dt \) are more muted for the \( SM_{05} \) MBE (Fig. 3c). For all \( dT/dt \) percentiles, the MBE has a small mean negative bias.
These analyses highlight an important feature and a unique aspect of the MDM approach that we applied in the present study, which is that none of the relationships for the different scenarios follow the mean diurnal cycles. The mean diurnal cycle has been used in many previous studies evaluating the HRRR (e.g., Lee et al. 2019; Patel et al. 2021; He et al. 2023; L23). Our findings reveal the importance of a weather regime–specific approach for model evaluation and suggests that, when the MDM is not evaluated for different near-surface atmospheric conditions, the MDM bulk statistics (the MBE in this context) provide a somewhat dubious answer. Here, the MDM bulk statistics complicate our ability to properly isolate the scenarios in which model deficiencies are most substantial. In the case of T, the largest MBEs occur under the largest \( \frac{dT}{dt} \). In these instances, the HRRR is likely not well representing land surface forcings during the early to mid-morning hours, which contributes to the large T biases. However, as noted above, the impacts of different observed \( \frac{dT}{dt} \) also affect the magnitude of the MBE for SW\(_d\). Furthermore, there are significant impacts on the SW\(_d\) MBE, with a larger positive MBE for small heating rates (Fig. 3c), and the MBE show a slightly mean negative bias for all \( \frac{dT}{dt} \).

2) INCOMING SHORTWAVE RADIATION

Previous studies have noted differences in the HRRR’s performance as a function of cloud cover (e.g., Min et al. 2021; L23). Accordingly, we specifically evaluate HRRRv4’s performance for different SW\(_d\) regimes (Figs. 3d-f). We find positive biases of nearly 200 W m\(^{-2}\) on days with the smallest SW\(_d\), and the HRRRv4 generally has a negative bias on days with the largest SW\(_d\). In contrast, the MBE resulting from averaging across all days was up to 100 W m\(^{-2}\) lower than the biases found on the cloudiest days. Furthermore, we note substantial impacts on the diurnal cycle of T MBE. On days with the smallest SW\(_d\), there is a positive T bias of about 3°C during the afternoon, and the MBE is about 2°C through the reminder of the day. In contrast, the HRRRv4 underestimates daytime T by about 2°C during the daytime on days with the largest SW\(_d\) but, interestingly, slightly overestimates T between about 0600 and 1200 UTC. As we found when evaluating the impact of \( \frac{dT}{dt} \) on the MBEs, different SW\(_d\) have a negligible impact on the SM\(_{05}\) MBE.

The larger positive biases on cloudy days (i.e., days with the smallest SW\(_d\)) are known from previous studies evaluating the HRRR and have been attributed to the model’s difficulty resolving subgrid-scale clouds (e.g., Lee et al. 2019; Min et al. 2021; He et al. 2023). The MDM differences that arise as a function of varying SW\(_d\), coupled with findings from section 4a(1), underscore the need for caution when calculating MDM statistics for model evaluation without any constraints present.

3) SOIL MOISTURE

As for SW\(_d\), previous studies also noted differences in HRRR’s performance as a function of different soil moisture regimes. Over CONUS, the HRRRv4 overestimates SM\(_{05}\) for dry soils and underestimates SM\(_{05}\) for wet soils. These discrepancies have been attributed to a possible issue with soil moisture conductivity in the HRRR, which was speculated by L23. Because soil moisture helps

![Fig. 5. Spatial variability in (a) SW\(_d\) MBE when \( \frac{dT}{dt} \) < 25th percentile and (b) SW\(_d\) MBE when \( \frac{dT}{dt} \) > 75th percentile. (c),(d) The spatial variability in SM\(_{05}\) MBE when \( \frac{dT}{dt} \) < 25th percentile and \( \frac{dT}{dt} \) > 75th percentile, respectively. Only values obtained between 1800 and 0000 UTC daily were used in these analyses.](image-url)
govern the partitioning of available energy into sensible and latent heat (e.g., Stull 1988), it is critical to differentiate model performance among different soil moisture regimes to provide additional insights into model performance. Our analyses here indicate that, whereas SM MBE across all cases exhibit only a small negative bias, soil moisture biases exhibit substantial variability across the different SM05 regimes (Fig. 3i). There is a positive bias of \(0.12 \text{ m}^3 \text{ m}^{-3}\) when the observed SM05 is small (i.e., 25th percentile) and negative bias of 0.16 m\(^3\) m\(^{-3}\) when the observed SM05 is large (i.e., 75th percentile). This finding is consistent with L23 investigating the relationship between the SM05 MBE and observed SM05 over CONUS. The \(T_d\) and \(SW_d\) MBes, however, show little sensitivity to the SM regime (Figs. 3g,h); there is a slightly positive nighttime \(T\) bias and slightly negative daytime \(\bar{T}\) bias which is consistent with the mean diurnal cycle across all days. Similarly, the \(SW_d\) MBE for all the different SM05 percentiles exhibit the same diurnal characteristics. The limited change in temperature may be related to the asymmetric responses of the near-surface atmosphere to soil moisture conditions. For example, Berg et al. (2014) noted that the ABL response to dry soil moisture conditions can lead to either warmer or cooler conditions, depending on the development of localized circulations (e.g., Stéfanon et al. 2014), clouds (e.g., Ek and Holtslag 2004; Lyons 2002), and precipitation (e.g., Findell and Eltahir 1997).

b. Spatial variability in biases

1) AIR TEMPERATURE

To further illustrate the impact of local changes in near-surface atmospheric conditions on HRRR biases, we consider the spatial variability among the USCRN sites in these biases and found substantial regional differences in MBEs (Fig. 4). The absolute values of the \(\bar{T}\) MBes are considerably less over CONUS when all cases are averaged together (Fig. 4a) than when differentiating between days for two extreme \(dT/dt\) regimes, i.e., days with \(dT/dt < 25\)th percentile (Fig. 4b) versus days with \(dT/dt > 75\)th percentile (Fig. 4c). In the latter two instances, there are markedly negative and markedly positive, respectively, \(T\) biases across CONUS. Averaging these negative and positive biases together to determine the MDM spatial variability for all cases, as shown in Fig. 4a, greatly reduces the MDM magnitude which is not necessarily a true representation of model performance. This finding underscores the importance of partitioning MBE computation by daytime heating regimes across the CONUS to allow for a more comprehensive, weather-specific MDM experiment.

Furthermore, when we evaluate the sensitivity of \(SW_d\) MBE and \(SM_{05}\) MBE to different percentiles of \(dT/dt\), we find that the HRRR generally has larger \(SW_d\) MBE in the subset of cases with \(dT/dt < 25\)th percentile than for the cases with \(dT/dt > 75\)th percentile. However, the \(SM_{05}\) MBE does not greatly vary as a function of the different \(dT/dt\) percentiles (Fig. 5). In general, soil moisture controls land-atmosphere feedback processes, with higher (lower) soil moisture yielding lower (higher) near-surface temperatures (e.g., Ek and Holtslag 2004; Pal and Haeffelin 2015; Pal et al. 2020). However, our findings suggest that changes in \(dT/dt\) regimes occurring under various weather conditions do not directly impact soil moisture forecasts. Nevertheless, whether the chain of feedback processes...
related to weather conditions and associated precipitation regimes (i.e., rain rate, type, and droplet size distributions), and consequently the variability of soil moisture and temperature, are impacted by different $dT/dt$ requires further investigation that is beyond the scope of the present study.

2) INCOMING SHORTWAVE RADIATION

Consistent with the results reported in section 4a(2), there is a slightly positive MBE (Fig. 4d) across CONUS. The bias is largest on the cloudiest days, i.e., when $SW_d < 25$th percentile (Fig. 4e) and exhibits little regional variability. In contrast and consistent with the mean diurnal MBE cycle (cf. Fig. 3e), the MBE for days with $SW_d > 75$th percentile is considerably smaller than days with the $SW_d < 25$th percentile (Fig. 4f). When we evaluate the sensitivity of $T$ MBE and $SW_d$ MBE to different percentiles of $SM_{05}$, we note considerable differences in the $T$ MBE for the different classifications of $SW_d$; there is a much smaller $T$ MBE for low $SW_d$ than for high $SW_d$, with many locations having MBE $> 3^\circ C$ (Fig. 6). As is the case for different classes of $dT/dt$, $SM_{05}$ MBE does not greatly vary as a function of different $SW_d$.

3) SOIL MOISTURE

The $SM_{05}$ MBE has a west-to-east gradient across all $SM_{05}$ percentiles, with positive MBEs across the drier western United States and negative MBEs across the wetter eastern United States (Fig. 4g), in agreement with L23. When soils are dry (i.e., $SM_{05} < 25$th percentile), the MBEs are typically positive across much of CONUS (Fig. 4h), whereas when soils are wet (i.e., $SM_{05} > 75$th percentile) MBEs are typically negative, with the exception of the very dry stations in the southwestern United States (Fig. 4i). These findings provide further evidence to support our hypothesis. In this particular scenario, contrasting MDMs under two different SM states (cf. Figs. 4h,i) are averaged out and thus are not represented in the bulk statistics (cf. Fig. 4g).

Finally, when we evaluate the sensitivity of $T$ MBE and $SW_d$ MBE to different percentiles of $SM_{05}$, we find that, when dry conditions are present (i.e., $SM_{05} < 25$th percentile), there are strong negative $T$ MBEs (i.e., $<-3^\circ C$) in the Rockies and positive MBEs across the eastern United States. The strongly negative $T$ MBEs in the Rockies are also present for wet conditions (i.e., $SM_{05} > 75$th percentile), whereas the $T$ MBEs are much smaller (Fig. 7). In contrast, there is little impact of the different $SM_{05}$ classifications on $SW_d$ MBE.

5. Summary, conclusions, and outlook

We demonstrated that weather regime–specific evaluations of NWP models are required to better distinguish cases when the models perform well versus when models perform poorly. We used one year of observations obtained from the USCRN coupled with HRRRv4 output and found substantial differences in model performance for the different near-surface atmospheric conditions. When averaged across all near-surface atmospheric conditions that we observed during the 1-yr study period, which is the typical approach used for evaluating weather forecasting.
models, the model biases tend to be rather small. However, the model biases greatly varied when distinguishing among different near-surface heating rates and radiative regimes. We found a negative $T$ MBE for small heating rates and positive $T$ MBE for larger heating rates, as well as substantial differences in $T$ and SW$_A$ MBE as a function of different radiative regimes.

Information about the aforementioned biases in operational forecasting models such as the HRRR and conducting additional similar evaluations of model performance is important for isolating the scenarios when model deficiencies are largest. The approach is arguably better than evaluating model performance using bulk statistics computed across a large range of different near-surface atmospheric conditions. Having knowledge about the conditional behavior of model biases provides insights into processes that are not well represented within the modeling system and are being pursued using a process-oriented verification approach described by, for example Turner et al. (2020). Doing so will permit future opportunities to evaluate NWP models under a myriad of other types of meteorological conditions. These conditions include, for example, nonsheared versus sheared ABLs (e.g., Fedorovich and Conzemius 2008), dry versus moist ABLs (e.g., Stull 1988), warm versus cold sectors of the ABL before and after frontal passages (e.g., Pal et al. 2021; Clark et al. 2022), etc. The use of a process-oriented verification approach will allow for targeted areas for improving the land surface and boundary layer parameterization schemes used within operational NWP models, like the HRRR and its successors (i.e., the RRFS). Finally, a regime-specific model evaluation, like the one developed in the present study, may similarly be used to help evaluate the efficacy of other geophysical models, including for example inverse carbon transport models, general circulation models, etc.

Acknowledgments. We thank NOAA/Air Resources Laboratory’s engineers for maintaining USCRN and our colleagues from NOAA/National Centers for Environmental Information for ensuring a high-quality dataset from USCRN. SP was partially supported by the NOAA Grant NA21OAR4590361. We also thank the two anonymous reviewers whose suggestions helped us to improve the technical and scientific content of the manuscript. Finally, we note that the results and conclusions of this study and views expressed herein are those of the authors and may not necessarily reflect the views of NOAA or the Department of Commerce.

Data availability statement. The USCRN datasets are archived by NOAA’s National Centers for Environmental Information and can be accessed at https://www.ncdc.noaa.gov/access/crn/qcdatasets.html. The HRRRv4 GRIB2 forecast output is available from the NOAA Open Data Dissemination program via Amazon Web Services’ archive, i.e. https://registry.opendata.aws/noaa-hrrr-pds/.

REFERENCES


