Benchmarking the Raw Model-Generated Background Forecast in Rapidly Updated Surface Temperature Analyses. Part I: Stations

THOMAS M. HAMILL

Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 7 February 2019, in final form 12 November 2019)

ABSTRACT

High-quality, high-resolution, hourly unbiased surface (2 m) temperature analyses are needed for many applications, including training and validation of statistical postprocessing applications. These temperature analyses are often generated through data assimilation procedures, whereby a background short-range gridded forecast is adjusted to newly available observations. Even with frequent updates to newly available observations, surface-temperature analysis errors and biases can be comparatively large relative to errors and biases of midtropospheric variables, especially over land, despite more near-surface in situ observations. Larger near-surface errors may have several causes, including biased background forecasts and the spatial heterogeneity of surface temperatures that results from subgrid-scale surface, vegetation, land-use, and terrain variations. Are biased raw background forecasts the predominant cause of surface temperature analysis errors? Part I of this two-part series describes a simple benchmark for evaluating the error characteristics of short-term (1 h) raw model background surface temperature forecasts. For stations with a relatively complete time series of data, it is possible to generate an hourly, diurnally, and seasonally dependent observation climatology at a station. The deviation of the current hour's temperature observation with respect to this hour's and Julian day's climatology is added to the climatology for the next hour. For contiguous U.S. stations in July 2015, the station benchmark was lower in error than interpolated 1-h high-resolution numerical predictions of surface temperature from NOAA's High-Resolution Rapid Refresh (HRRR) system, although not including full postprocessing. For August 2018, 1-h HRRR forecasts were much improved when tested against the station benchmark.

1. Introduction

Many weather and climate applications require high-quality hourly surface (2 m) real-time temperature analyses and retrospective analyses (reanalyses). For example, an accurate retrospective time series of surface temperature analyses on a high-resolution grid may be used to provide the analyzed training data for the statistical postprocessing of surface temperature forecasts, such as in Flowerdew (2014). Other applications include accurate model initialization, the diagnosis of climate and weather variations and trends, the validation of surface-temperature forecasts from numerical weather prediction guidance, and situational awareness of current conditions. Users seek analyses with low error and bias as well as realistic spatial and temporal detail, including smaller diurnal temperature ranges near water bodies and elevation-dependent and terrain slope-dependent temperature variability in mountainous regions. If the analyses are biased, have large errors, or have insufficient spatial and temporal detail, they may be unsuitable for these applications. The U.S. National Oceanic and Atmospheric Administration (NOAA) maintains a system that produces such hourly "analyses of record:" the Real-Time Mesoscale Analysis (RTMA) system of de Pondeca et al. (2011).

How accurate and unbiased are the current hourly surface temperature analyses? Commonly these analyses are generated with data assimilation algorithms, whereby a first-guess "background" forecast is adjusted to newly available observations. Data assimilation algorithms commonly are formulated under the assumption that the background forecasts are unbiased (e.g., Daley 1991, chapter 4). This assumption should be checked rather than taken as a given. For example, significant systematic analysis increments (analysis minus background) were identified in the raw NOAA High-Resolution Rapid Refresh (HRRR; Benjamin et al. 2016) short-range surface temperature forecasts during July 2015 that are used in the generation of the background

Corresponding author: Dr. Thomas M. Hamill, tom.hamill@noaa.gov

DOI: 10.1175/MWR-D-19-0027.1

forecasts for the RTMA analysis procedure. The HRRR is a fully nonhydrostatic forecast modeling system that utilizes a modern land surface model (LSM) for estimating surface-energy fluxes and diagnosing surface temperature. Time-averaged analysis increments (analysis minus forecast) for 0000 UTC during July 2015 and August 2018 are shown in Figs. 1a and 1b, respectively. Warm biases are evident in much of the eastern and central United States during July 2015. Raw background biases were substantially reduced by August 2018, primarily as a result of more recent model and assimilation changes outlined in Benjamin et al. (2016, their Table 8). Of course, biased raw surface-temperature background forecasts are not unique to the HRRR system; they can be observed in practically all data assimilation systems.

While systematic discrepancies between surface temperature observations and raw model forecasts at particular stations may be attributable in part to differences in grid elevation between the model grid and the observation, there are other causes that may relate to underlying model imperfections. If there are systematic errors in ground-heat flux or the soil-water budget, perhaps due to a mis-estimation of precipitation, soil moisture, or soil texture or a deficiency in the land surface model, these may accumulate in the absence of a corrective soil-state data assimilation procedure. The biased soil temperature states may in turn result in biased estimates of fluxes of thermal energy and moisture between the ground and the air above it, affecting surface temperature. There may also be prediction system issues that are traced back to model deficiencies above the ground surface. Perhaps surface downward solar radiation is systematically misestimated because of an inappropriate forecast of cloud cover and optical depth, or perhaps there are systematic errors in the surface-layer or boundary layer physical parameterizations that result in misestimations of vertical mixing.

Could bias in the raw background state be removed prior to the assimilation step? With regards to differences that originate from fixed differences between the station and the grid elevation, the RTMA as well as the fully cycled HRRR data assimilation procedures incorporate adjustments to the raw model background for differences in elevation. The assimilation procedures for these two systems are shown in Fig. 2. In the RTMA, lapse rates from the HRRR forecast are used in conjunction with differences between the finer-scale RTMA grid resolution and the coarser HRRR grid resolution to define a modified background state consistent with the RTMA grid elevation (de Pondeca et al. 2011). This vertical downscaling is described more in Benjamin et al. (2007) and is also similar to that described by Huld and Pascua (2015). In grid cells with fractional water coverage,

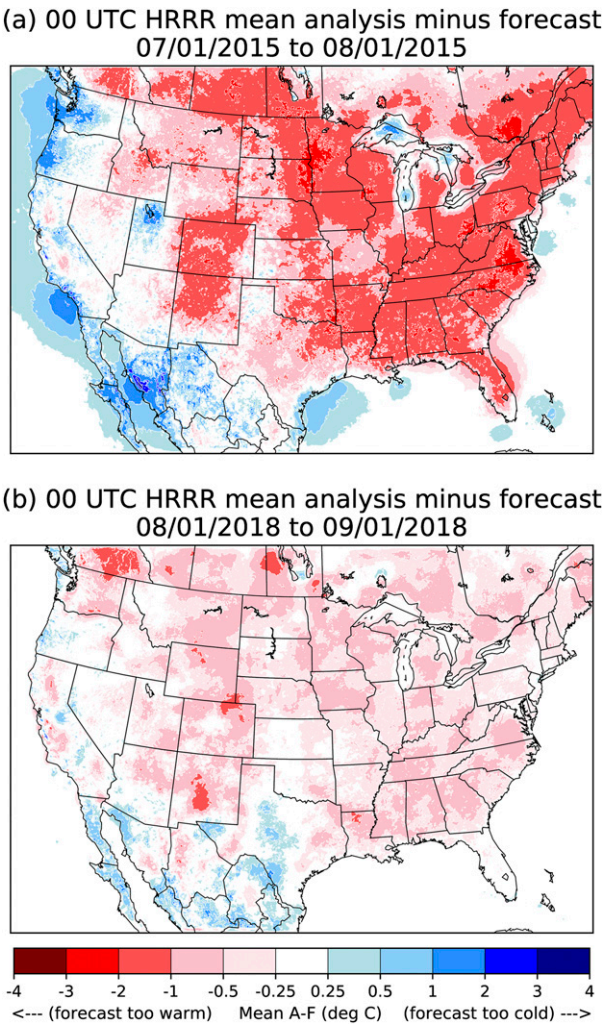

FIG. 1. Illustration of potential systematic errors in 1-h HRRR forecasts and their change from 2015 to 2018. Mean analysis increments (analysis minus forecast) over CONUS are shown for the 1-h forecasts ending at 0000 UTC during (a) July 2015 and (b) August 2018.

there is also a procedure for choosing the background state at an adjacent grid point if the observation minus background is smaller at that point. The production of 2-m temperature analyses in the cycled HRRR 3D data assimilation for model initialization is slightly different, in that all observations, surface and other, are assimilated. Adjustment for elevation differences are performed using differences between the HRRR model grid and the observation elevation (Benjamin et al. 2016, their section 2a).

There is also literature that discusses the more generic problem of larger-scale biases in raw background forecasts and possible corrective approaches (e.g., Dee and Da Silva 1998; Dee 2005; Baek et al. 2006; Lei and Hacker 2015; Lorente-Plazas and Hacker 2017). Still, removing background bias is especially challenging for

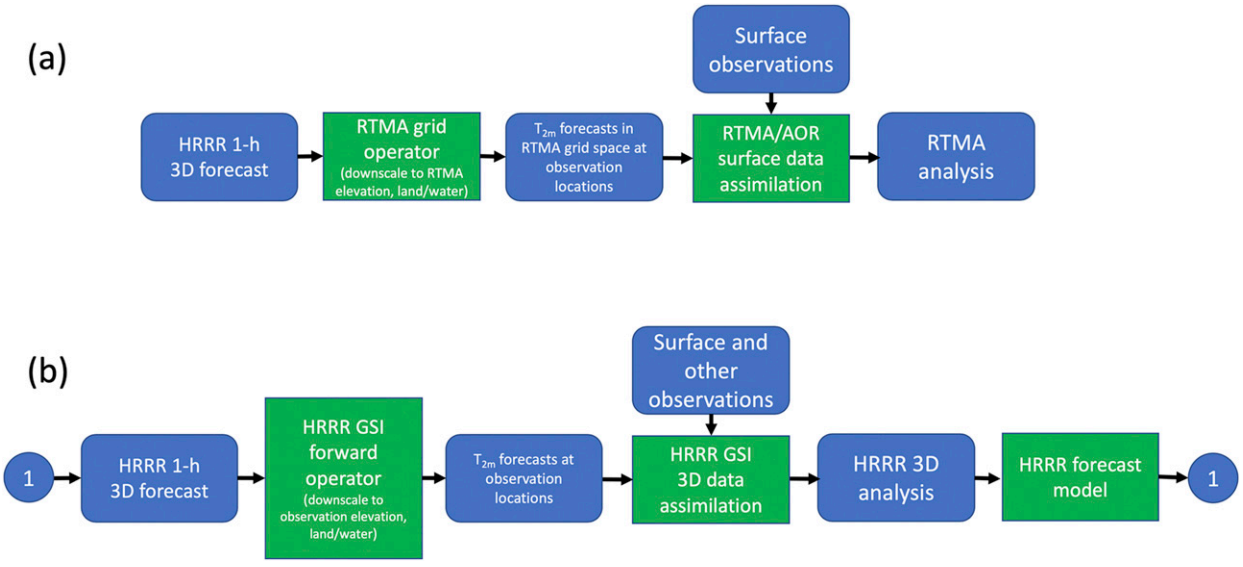

FIG. 2. Schematic of the assimilation procedures for the (a) RTMA and (b) HRRR systems. The RTMA system is not cycled; i.e., the analyses are not used to initialize the subsequent forecast. Elevation adjustments are made between HRRR and RTMA grids to the background forecast in the RTMA analysis procedure and between HRRR and station elevations in the HRRR analysis procedure.

surface temperature given its spatial heterogeneity. Whereas one might pool samples of, say, stratospheric temperatures across many locations to increase the sample size for bias estimation, this may be counterproductive with surface temperatures.

It would be helpful to understand whether raw model background surface-temperature biases in rapidly updated surface data assimilations are a minor nuisance or a major problem, in this case before HRRR adjustments. Best et al. (2015, hereinafter B15) discussed the use of *benchmarks* to provide baseline performance expectations for a system; the benchmark might be a previous model version or a simple statistical model. In B15, the hypothesis was that sophisticated land surface models would be able to provide more accurate forecasts of sensible and latent heat fluxes when compared to simple statistical-model benchmarks. Fluxes were benchmarked at a range of sites with high-quality observed data (e.g., time series of surface temperature and humidity, insolation, and soil temperature and moisture). The hypothesis was not confirmed; surface sensible and latent-heat flux estimates from LSMs were generally less accurate than a statistical benchmark based on 3 simple predictors, the surface temperature, relative humidity, and downward solar-radiation at the surface. As surface temperature is strongly related to the partitioning of fluxes at the earth's surface, this motivated the development of a simple statistical benchmark of surface temperature to compare with 1-h forecasts from a sophisticated prediction system. Given the results of B15, the hypothesis to be tested that a station-based statistical 1-h forecast of surface temperature, hereinafter called the "station benchmark" will provide a challenging reference standard for short-range surface temperature raw forecasts from a high-resolution numerical model.

This article is inspired by B15 and is the first of a two-part series. This first part discusses the development of the station benchmark for raw numerical HRRR background forecasts of surface temperature for the contiguous United States (CONUS). Evaluation of raw HRRR model data is a chosen simplification that neglects evaluation that includes the temperature adjustments for elevation in the RTMA and the HRRR. Arguably the simplified data processing does illuminate the underlying error characteristic of the raw model guidance. The article intends both to further demonstrate the relevance of benchmarking and to demonstrate the substantial challenges in developing numerical weather prediction capable of ameliorating short-term forecast bias. As an ancillary research result, the article will demonstrate the utility of a careful definition of a diurnally dependent climatology.

Admittedly, the evaluation of raw gridded model guidance relative to a station benchmark may be misleading. The observation site for a station benchmark may reflect conditions unique to that particular location rather than the surrounding gridbox mean that the prediction system represents. Still, if a station benchmark does not set a competitive standard for the model guidance, then one would not expect a statistically generated and rigorously cross-validated *gridded* benchmark to be competitive with model guidance. As we will see,

the station benchmark does provide a competitive reference standard, which motivates further development of the gridded benchmark discussed in the second part of this series, Hamill and Scheuerer (2020, hereinafter Part II).

The remainder of this first article is organized as follows. Section 2 describes the data used in this experiment and the methods for evaluation of the raw forecast and station benchmark. Section 3 describes the numerical procedure used to generate the station benchmark. Sections 4a and 4b provide results with regard to the verification of the benchmarks, and section 5 provides a discussion and conclusions.

2. Data and evaluation methods used in this experiment

The observation dataset used in this experiment was the National Center for Atmospheric Research dataset 472.0, an archive of quality-controlled hourly surface observations over North America. Data were originally synthesized and quality controlled at the U.S. National Weather Service Meteorological Development Laboratory. These data are available online (https://rda.ucar.edu/datasets/ds472.0/). Surface temperatures were used for the period 0000 UTC 1 January 2004–2300 UTC 28 February 2019. The author chose to further limit use of surface temperatures in this dataset to only those observation sites for which data were available at 97% or more of the hours, days, and years in the analysis period. This observation availability cutoff was made based on the importance of an accurate estimation of the climatology to this procedure. With this availability criterion, 1118 station locations were available in the area of study, the CONUS.

When comparing the benchmarking procedure with numerical forecasts, for the July 2015 data, 1-h forecasts of background surface temperatures were extracted from version 1 of the operational High Resolution Rapid Refresh (HRRR) limited-area prediction system described in Benjamin et al. (2016). The raw forecast value at the ~3-km^2 grid box nearest the station was used, a simplification of the process described in Fig. 2. For the 2018 data, the operational version-3 HRRR predictions were extracted. The HRRR system generates hourly analyses and numerical forecast guidance to +15-h lead time. It is used for many applications at the National Weather Service (NWS), including severe weather prediction, short-term precipitation prediction, and aviation applications. The underlying prediction system is the Weather Research and Forecasting (WRF) Advanced Research WRF (ARW), with a 3D-ensemble-variational data assimilation system. See Benjamin et al. (2016) for more details. Comparative validation of station benchmark and HRRR forecasts was limited to July 2015 and August 2018—a limitation of this study.

July 2015 HRRR forecasts were sometimes unavailable—in particular, 1-h forecasts initialized on the date/times (all UTC) 1500 1 July, 1000 2 July, 0500 and 1400 3 July, 1400 5 July, 1200 and 2100 6 July, 0800 8 July, 0300 10 July, 0800 and 1600 11 July, 1400 and 2100 18 July, 1300 22 July, 1100 23 July, 1300 UTC 26 July, and 2100 UTC 28 July. This set represents 17 of the 744 analysis times, or approximately 2.3%. The validation of both the HRRR and the station benchmark did not include these data. No data were missing in 2018.

Standard methods of evaluation of deterministic forecasts were used, including root-mean-square error (RMSE), mean absolute error (MAE), and bias, all following standard definitions in Wilks (2011). 5th- and 95th-percentile confidence intervals of a distribution consistent with the null hypothesis of no differences are provided on the comparative plot of errors from the two systems, recentered on the benchmark forecast errors. The confidence intervals were determined through a paired block bootstrap algorithm following Hamill (1999), assuming error statistics were independent from one day to the next (Hamill 1999).

3. Methods used in the generation of the statistical benchmark

To determine a current hour's deviation from climatology at a particular station, an accurate estimate of that climatology is needed. In this application, the climatology was estimated to be a function of the hour of the day (which permitted diurnal dependence) and of the Julian day of the year (which permitted seasonal dependence). With such estimates, it was straightforward to generate the station benchmark for 1-h dynamical surface-temperature forecasts; the current hour's observed anomaly with respect to that hour's climatology was determined. This anomaly was added to the next hour's climatology to generate the 1-h station benchmark.

Figures 3 and 4 illustrate the procedure for generating the seasonally and diurnally dependent temperature climatology for a particular station, in this case the airport at Albany, New York (KALB). Figure 3 shows the 0000 UTC observations at KALB (dots) as a function of the Julian day. Plotted over the top of these is a cubic-spline fit estimate (Press et al. 1992, section 3.3) of the mean temperature as a function of the Julian day. To generate this curve, data were repeated below Julian day 1 and after Julian day 365, and the cubic-spline procedure was applied using eight knots equally spaced through the calendar year. The choice of eight knots was based on trial and error for what appeared

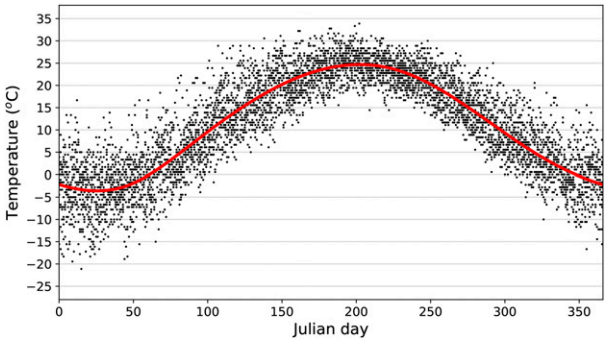

FIG. 3. Illustration of the cubic-spline fitting procedure to determine a climatology for a particular hour of the day for KALB. Surface temperatures (dots) are shown for the period from 0000 UTC 1 Jan 2004 to 2300 UTC 28 Feb 2019. The fitted spline-curve estimate of the climatology for this period is shown with the red curve, estimated as a function of Julian day. To aid clarity of presentation, only every fifth sample from the time series was plotted.

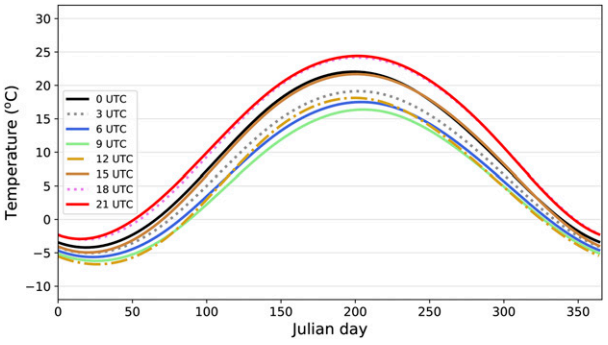

FIG. 4. Spline-fitted estimates of the climatology at KALB for every 3 h over the diurnal cycle and throughout the year.

to provide a reasonable, smooth fit to the data. Figure 4 next shows the cubic-spline-fitted yearly climatologies at KALB every third hour over the diurnal cycle. The diurnal temperature range was smallest in the boreal midwinter and largest in midsummer. Minimum temperatures were most commonly closer to 1200 UTC in midwinter, but with the earlier sunrise were nearer to 0900 UTC during the summer. These climatologies were generated for each of the 1118 CONUS stations and for every hour of the day, cross validated, so that when a data analysis was performed for 2004 the climatology was defined with the 2005–19 data. Possible systematic changes in climate from anthropogenic global warming were not considered in the definition of the climatology.

With the climatology defined, it was straightforward to evaluate the potential validity of this persistence of the deviation from climatology as a station benchmark. Figure 5 provides scatterplots of this station temperature anomaly relative to the temperature anomaly in the previous hour for eight chosen hours through the diurnal cycle, again for July at KALB. There were uniformly large 1-h-lag Pearson correlations of the anomalies at all hours, modest RMS differences on the order of 1°C, and an evident lack of bias. The procedure for generating the station benchmark appears to be rigorous across the diurnal cycle for this station. Although not shown, these general characteristics were confirmed when the analysis was repeated at other stations and other times of the year.

Figure 6 illustrates both the procedure for generating these deviations from climatology. Figure 6a shows both the hourly and Julian day-dependent climatology (thick blue line) and the hourly time series of observations (red line) for 1–15 July 2015. The observed deviations from the climatology are plotted in Fig. 6b. Deviations exhibited a modest autocorrelation, which may be due to synoptic-scale variability on a time scale of a week and the persistence from one hour to the next of environmental conditions that affect surface temperature such as cloudiness or soil-moisture and soil-temperature anomalies.

Given the autocorrelations, a more complicated benchmark was developed with linear-regression corrections to the persistence of the anomaly. Predictors included the relative humidity and temperature trends above the surface from a forecast model. This further decreased the error of the station benchmark by approximately 5%–10%. However, this more complicated model was not used as a benchmark–the focus hereinafter is on the simplest of procedures, direct persistence of the anomaly from climatology.

4. Results

a. Verification of the benchmark for 2004–19

Before comparing the benchmark with the HRRR data, a figure is provided to illustrate that the benchmark has modest error and bias when validated at many different stations in different climatological regimes and when validated over yearly and diurnal cycles. Figure 7 shows that CONUS-averaged benchmark RMSE, MAE, and bias were all modest over the yearly and diurnal cycles. Errors tended to be largest during the time of maximum heating, which was later in the day during the wintertime. It is during the period of maximum heating when the accuracy of the partitioning of the downward solar and thermal energy into surface sensible heating, latent heating, and ground heat flux has the most consequence. If the partitioning is misestimated, perhaps as a result of errors in the analysis of soil moisture or the forecast of cloud cover, then the rate of warming will in turn be misestimated.

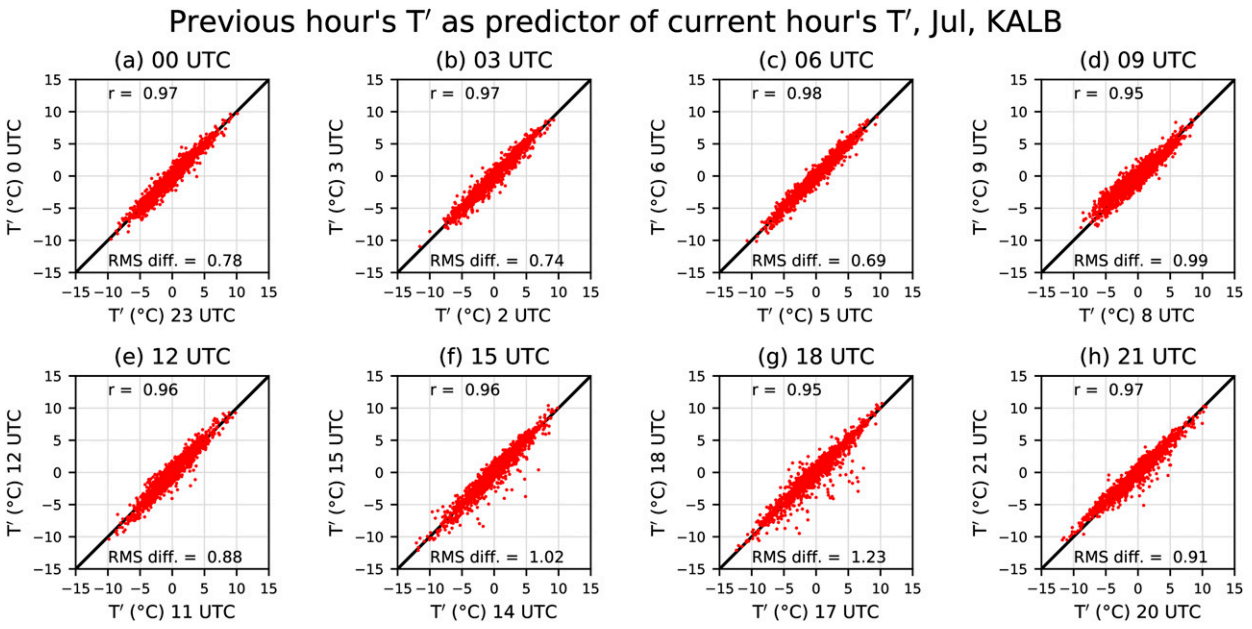

FIG. 5. Scatterplots of deviation T' from climatology at station KALB for a particular hour of the day (ordinate) vs the deviation for the previous hour (abscissa). Data are collected for dates in July from 2004 to 2018. Here, (a)–(h) show the relationships every 3 h over the diurnal cycle.

Another characteristic shown in Fig. 6c is that the bias is consistently near zero for all months and for all hours across the diurnal cycle.

b. *Verification of HRRR 1-h surface temperature forecasts against the benchmark*

Figure 8 synthesizes the comparative verification of the HRRR forecasts with the station benchmark during July 2015 (Figs. 8a,c,e) and August 2018 (Figs. 8b,d,f). During July 2015, the station benchmark was statistically significantly lower in RMSE and MAE, roughly slightly less than a factor of two over the diurnal cycle. The station-based benchmark was unbiased, while the HRRR system was commonly too warm during the daytime hours. This result was consistent with the averaged analysis increments from cycled data assimilation previously shown in Fig. 1. August 2018 benchmark errors were very similar to July 2015 benchmark errors, but the August 2018 HRRR errors were notably reduced relative to the July 2015 values. Benjamin et al. (2016; Fig. 11) shows similar error and bias reductions in 12-h forecasts.

Figure 9 provides a plot of RMSE errors at each station location. Again, note that the August 2018 HRRR RMSEs (Fig. 9b) were notably reduced relative to those in July 2015 (Fig. 9a). The July 2015 HRRR errors were largest in the upper Great Plains and at selected locations in the mountainous western United States. Figures 8c,d provide the respective maps of station benchmark RMSEs, which are generally uniformly low, though with a few scattered points in the Rocky Mountains with higher errors.

Figure 10 presents spatial maps of HRRR and station benchmark biases for the two summer months. The station benchmark biases are generally lower, and the HRRR bias is markedly reduced in August 2018 relative to its July 2015 values. Note that the pattern of July 2015 HRRR bias strongly resembles the time-mean analysis increment shown in Fig. 1a. Figure 11 provides a scatterplot of the station benchmark RMSE (Fig. 11a) and bias (Fig. 11c) against the HRRR at 0000 UTC in July 2015 and August 2018 (Figs. 11b and 11d). Again, the majority of station benchmark forecasts were lower in error for July 2015, though the proportion of benchmark RMSEs that were lower were reduced in August 2018. It is possible that if the plot were generated utilizing HRRR background forecasts after their output grid elevation correction, the HRRR errors would be reduced. Station benchmark biases averaged to near zero and were mostly confined between −1° and 1°C. HRRR forecasts in July 2015 were much more commonly too warm than too cold, although the HRRR biases were notably reduced in August 2018. Overall, the results of the station benchmark provide enough evidence to confirm the original hypothesis, that a reasonable statistical benchmark is capable of improving upon raw dynamical short-term forecasts of surface temperature

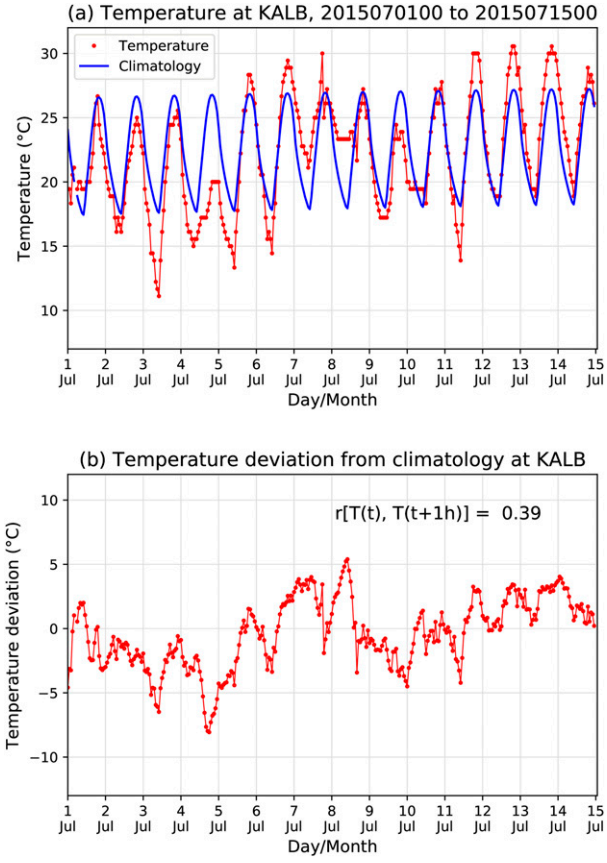

FIG. 6. Illustration of the process for development of the hourly benchmark: (a) hourly time series of surface temperature and the surface temperature climatology at KALB and (b) deviations of the hourly surface temperature from its climatology. The 1-h lagged Pearson autocorrelation is also indicated in (b) for the data. The vertical rules of the background grid represent 0000 UTC each day.

without postprocessing, although of course the station benchmark omits errors of representativeness.

5. Discussion and conclusions

In this article a simple procedure was developed to produce a benchmark for the evaluation of 1-h forecasts of surface temperature from a numerical weather prediction system. Such forecasts are commonly used as the background in hourly "rapid update" data assimilation. The procedure began with the development of a climatology for the station that varied with the Julian day and hour of the day. Observed deviations from the current hour's climatology were added to the next hour's climatology to produce the 1-h forecast station benchmark. This procedure was used during July 2015 and August 2018 to benchmark raw 1-h forecasts from the HRRR system. The HRRR data used here did not include the

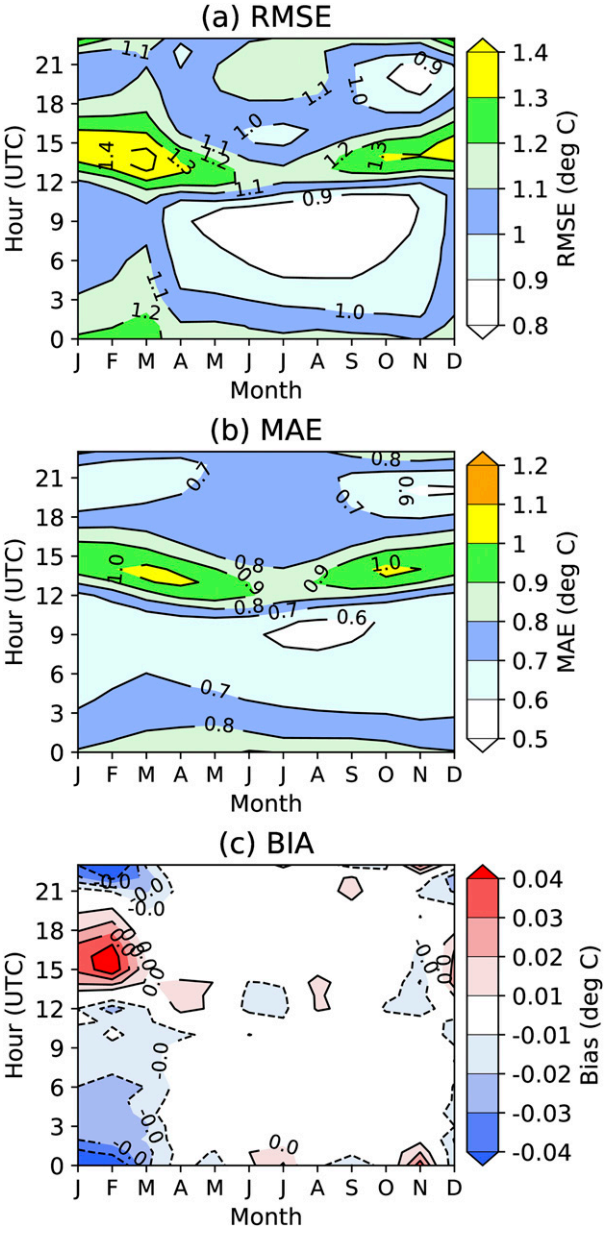

FIG. 7. (a) Root-mean-square error, (b) mean absolute error, and (c) bias for 1-h surface-temperature forecasts from the station benchmark for stations in the CONUS. The data span 1 Jan 2004–28 Feb 2019. Errors were plotted as a function of the month of the year (abscissa) and the initialization time for the 1-h forecast (ordinate).

postprocessing used in the RTMA system that includes a vertical interpolation to elevation differences between the HRRR and the RTMA grid. The station benchmark had statistically significantly lower errors and biases than the raw HRRR background forecasts, though the forecasts showed a notable improvement from 2015 to 2018. An admitted limitation of the study was that

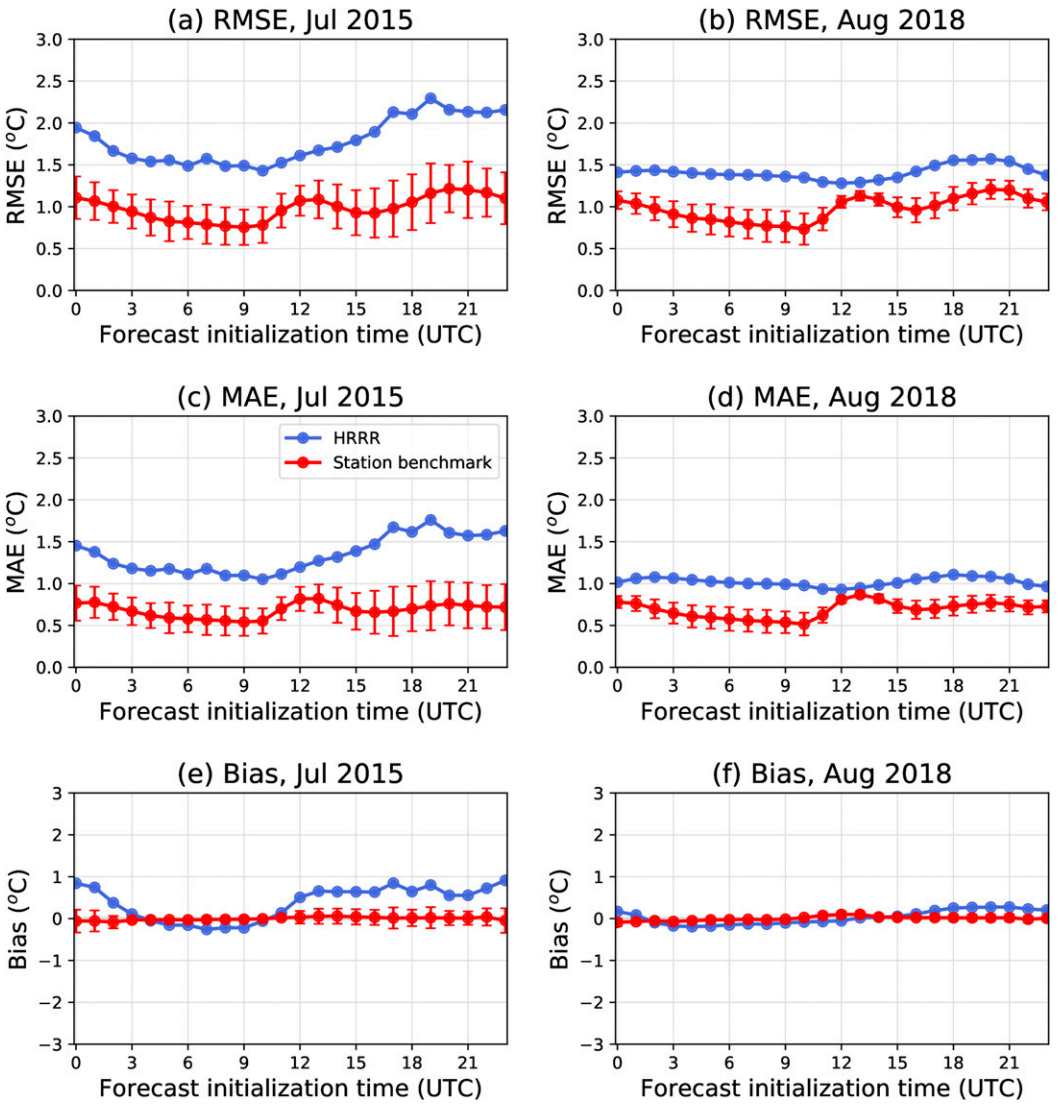

FIG. 8. The 1-h surface temperature forecast-error statistics for CONUS HRRR forecasts interpolated to stations and for the station benchmark at those stations: (a) root-mean-square error for July 2015, (b) RMSE for August 2018, (c) mean absolute error for July 2015, (d) MAE for August 2018, (e) bias for July 2015, and (f) bias for August 2018. Error bars are recentered around the station benchmark and represent the 5th and 95th percentiles from a paired block bootstrap distribution consistent with the null hypothesis of no differences in mean.

cool-season forecasts were not evaluated; this would be rational to examine in future research.

The use of a station-based benchmark of a numerical weather prediction is of course problematic for the reasons discussed earlier. The model by design estimates a gridbox-averaged value, and its initialization incorporates the effect of other nearby observations. A more realistic benchmark would thus be a gridded statistical benchmark, ideally one where the validation of a 1-h forecast at a station location did not use the information from that station during the previous hour. That is precisely what part 2 of this article will construct and evaluate.

Despite the reservations about the validation against station data, this simple benchmark of surface temperatures, like the B15 benchmark of surface fluxes, is thought provoking, especially for model development groups. The perceived advantage of using a numerical weather prediction system in forecasting future states is of course its ability to predict changes in air masses and associated weather conditions, demonstrated in innumerable studies. The station-based results suggest that the advantages in predicting very short-term changes in weather conditions is counteracted to some extent by the substantial numerical challenges involved in successfully

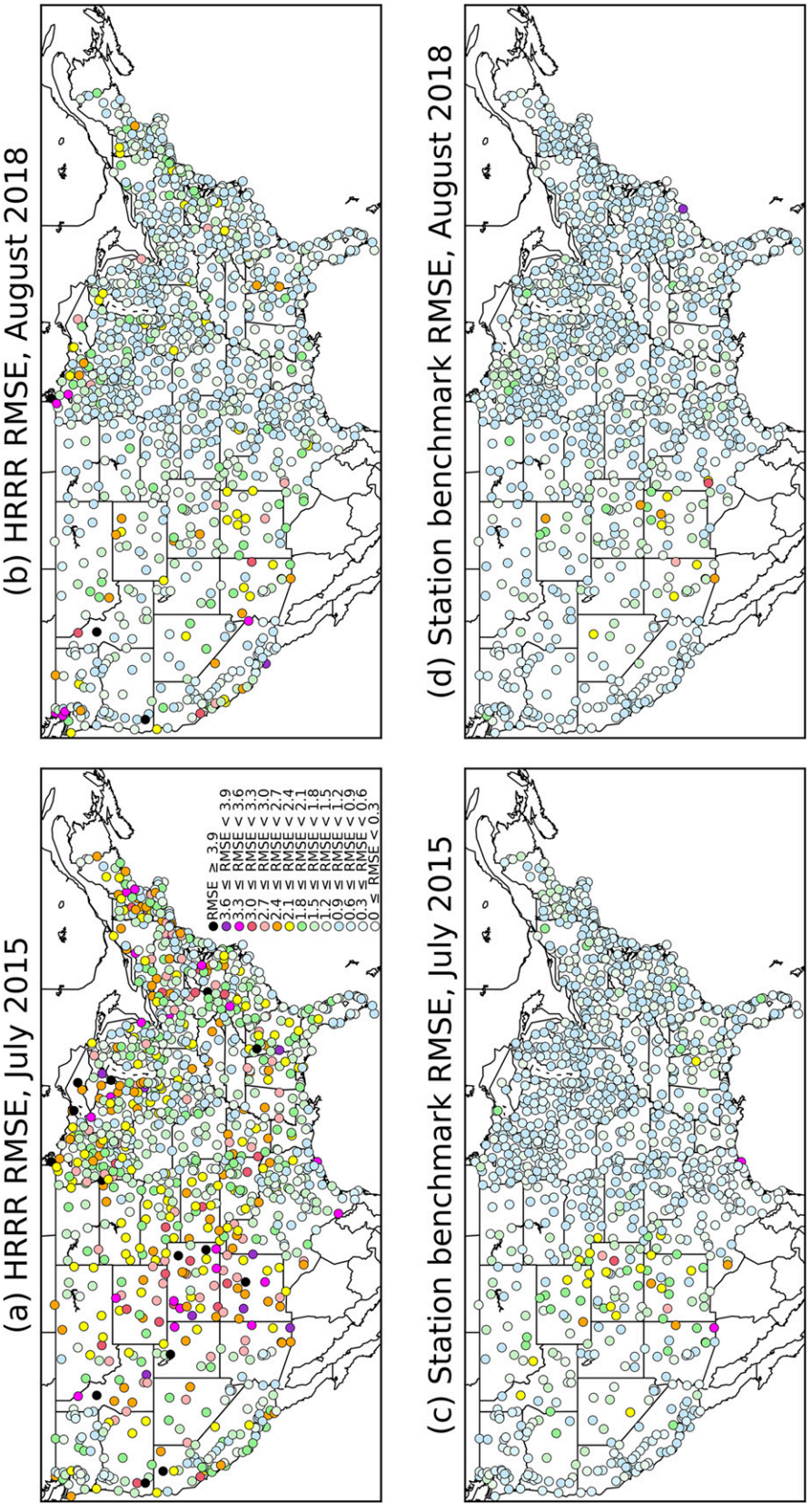

FIG. 9. RMSEs of 1-h forecasts initialized at 0000 UTC for (a),(b) HRRR model forecasts and (c),(d) station benchmarks for (left) July 2015 and (right) August 2018.

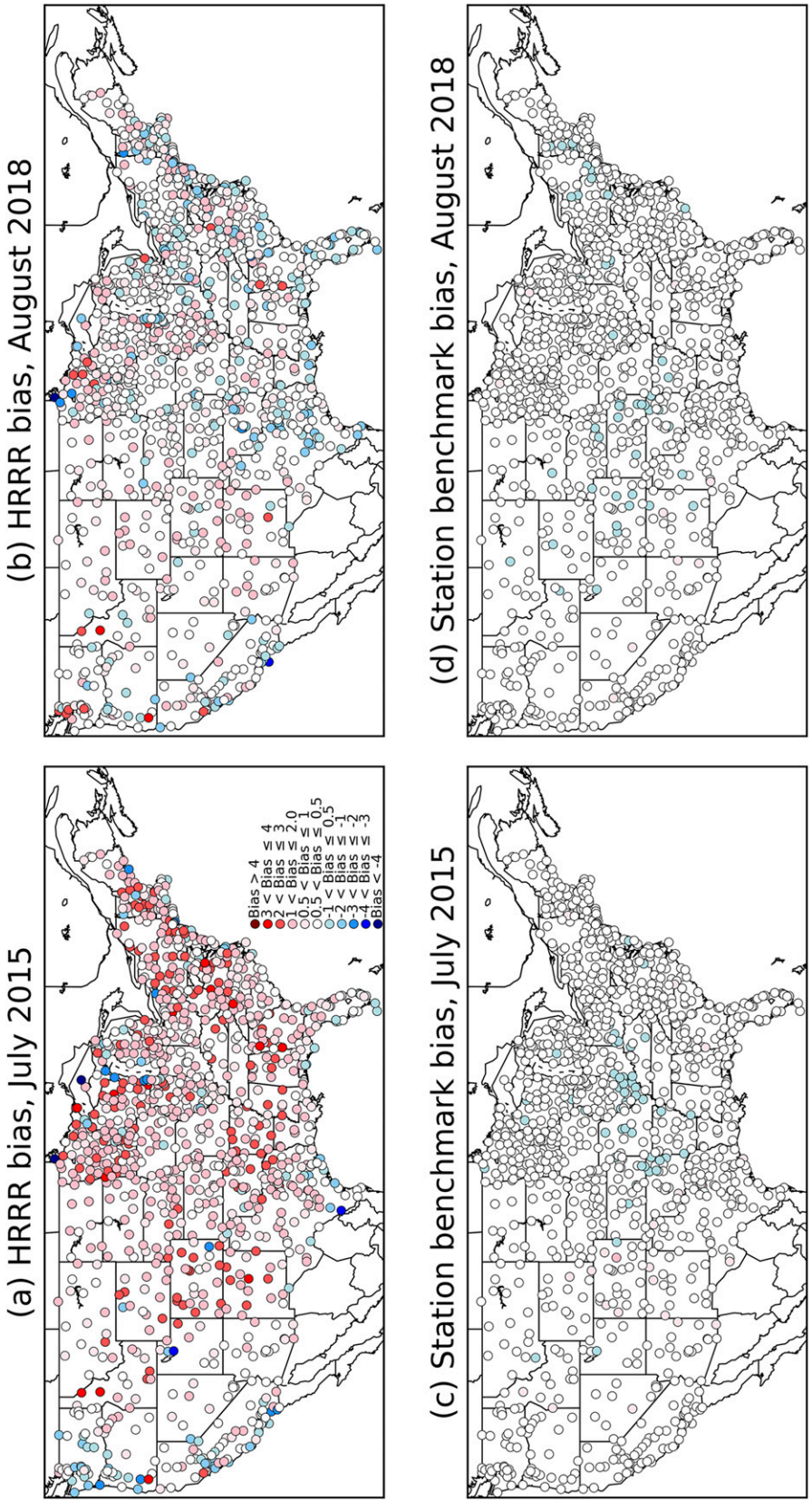

FIG. 10. As in Fig. 9, but for biases.

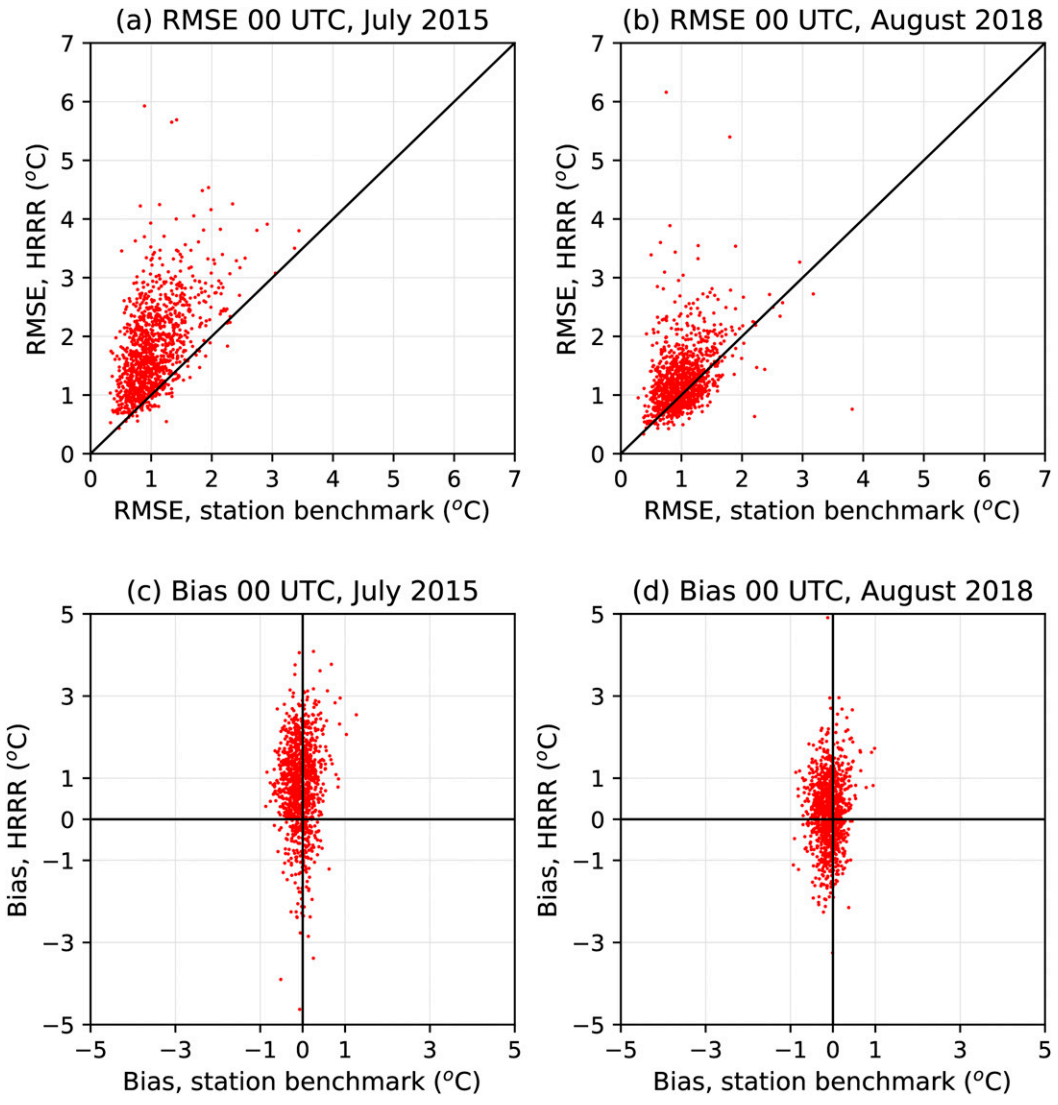

FIG. 11. Scatterplot of errors and bias for HRRR vs station benchmark, showing 1-h forecast (a),(b) RMSE and (c),(d) bias for (left) July 2015 and (right) August 2018.

analyzing and predicting the diurnal evolution of the surface state. These challenges include vertical interpolation and potential systematic errors in the atmospheric forecasts of solar radiation or mixing of winds near the ground. They may also include biased initial estimates of the soil state (temperature, moisture, or snow cover) and a suboptimal representation of the physical processes that govern the interaction of the land with the atmosphere. It remains to be seen if the lower error and bias in the station benchmark is preserved in a more apples-to-apples comparison of the raw HRRR background with a gridded benchmark.

The reader is now directed to Part II, which discusses the development of a statistical procedure for generating a 1-h gridded forecast of surface temperatures over land and the comparative evaluation of these temperatures relative to the HRRR guidance and then discusses the implications of the results from the two articles.

Acknowledgments. This work was supported partially by ESRL Physical Sciences Division base funding provided by NOAA's Office of Oceanic and Atmospheric Research and partially by U.S. NWS Office of Science and Technology Integration (through the Meteorological Development Laboratory) Projects R8MWQML-P00 and T8MWQML.P00. Eric James of the ESRL Global Systems Division is thanked for assisting with the extraction of the HRRR data from the tape archive. Stan Benjamin and two other anonymous scientists are thanked for their constructive peer reviews.

REFERENCES

Baek, S.-J., B. R. Hunt, E. Kalnay, E. Ott, and I. Szunyogh, 2006: Local ensemble Kalman filtering in the presence of model bias. *Tellus*, **58A**, 293–306, https://doi.org/10.1111/j.1600-0870.2006.00178.x.

Benjamin, S. G., J. M. Brown, G. Manikin, and G. Mann, 2007: The RTMA background—Hourly downscaling of RUC data to 5-km detail. *23rd Conf. on IIPS*, San Antonio, TX, Amer. Meteor. Soc., P1.11, https://ams.confex.com/ams/pdfpapers/119134.pdf.

——, 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.

Best, M. J., and Coauthors, 2015: The plumbing of land surface models: Benchmarking model performance. *J. Hydrometeor.*, **16**, 1425–1442, https://doi.org/10.1175/JHM-D-14-0158.1.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 457 pp.

De Pondeca, M., and Coauthors, 2011: The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: Current status and development. *Wea. Forecasting*, **26**, 593–612, https://doi.org/10.1175/WAF-D-10-05037.1.

Dee, D. P., 2005: Bias and data assimilation. *Quart. J. Roy. Meteor. Soc.*, **131**, 3323–3343, https://doi.org/10.1256/qj.05.137.

——, and A. M. Da Silva, 1998: Data assimilation in the presence of forecast bias. *Quart. J. Roy. Meteor. Soc.*, **124**, 269–295, https://doi.org/10.1002/qj.49712454512.

Flowerdew, J., 2014: Calibrating ensemble reliability whilst preserving spatial structure. *Tellus*, **66A**, 22662, https://doi.org/10.3402/tellusa.v66.22662.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, and M. Scheuerer, 2020: Benchmarking the raw model-generated background forecast in rapidly updated surface temperature analyses. Part II: Gridded benchmark. *Mon. Wea. Rev.*, **148**, 701–717, https://doi.org/10.1175/MWR-D-19-0028.1.

Huld, T., and I. P. Pascua, 2015: Spatial downscaling of 2-meter air temperature using operational forecast data. *Energies*, **8**, 2381–2411, https://doi.org/10.3390/en8042381.

Lei, L., and J. P. Hacker, 2015: Nudging, ensemble, and nudging ensembles for data assimilation in the presence of model error. *Mon. Wea. Rev.*, **143**, 2600–2610, https://doi.org/10.1175/MWR-D-14-00295.1.

Lorente-Plazas, R., and J. P. Hacker, 2017: Observation and model bias estimation in the presence of either or both sources of error. *Mon. Wea. Rev.*, **145**, 2683–2696, https://doi.org/10.1175/MWR-D-16-0273.1.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in FORTRAN*. 2nd ed. Cambridge University Press, 963 pp.

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