The Diurnal Cycle of Temperature Errors in the Operational Global Forecast System (GFS)

Ronak N. Patel[1], Sandra E. Yuter[1], Matthew A. Miller[1], Spencer R. Rhodes[1], Lily Bain[1], Toby Peele[1]

[1]Department of Marine, Earth, and Atmospheric Sciences, North Carolina State University, Raleigh, NC

Key Points:

- NOAA's GFS model struggles to adequately represent the diurnal cycle of temperatures under observed conditions of $\leq 50\%$ cloud cover
- NOAA's HRRR model, which uses a different parameterization suite, does not have a strong diurnal cycle of errors under the same conditions
- Examination of errors by similar weather conditions helps to constrain the portion of model physics that can yield larger forecast errors

Corresponding author: Sandra Yuter, `seyuter@ncsu.edu`

Abstract

Forecasts from NOAA's Global Forecast System (GFS) and the High-Resolution Rapid Refresh (HRRR) weather models are matched to surface observations for the winter season of November 2019 – March 2020 at 210 airports across the United States. The 2-meter temperature errors, conditioned on observed weather conditions such as cloud cover amount and wind speed, are used to determine the nature of systematic model biases. We observe a strong diurnal cycle in 2-meter temperature errors in the GFS model in conditions with $\leq 50\%$ cloud cover, with a 1° C warm bias at night and a 2° C cold bias during the day. The HRRR model, which uses a different set of physical parameterizations, does not have a clear diurnal cycle in errors under the same conditions. These results highlight the utility of weather-conditional comparisons across the diurnal cycle to diagnose sources of model weaknesses and to target model improvements.

Plain Language Summary

We evaluate the output of weather forecast models compared to observations at 210 airports across the United States during the November 2019 to March 2020 winter season. We focus on near-surface air temperature errors in the Global Forecast System (GFS) and High-Resolution Rapid Refresh (HRRR) weather models for different times of day and for subsets of observed weather conditions. The GFS is 1° C too warm at night and 2° C too cold during the day in conditions with less than 50% cloud cover. The daily high and low temperatures have smaller errors in the HRRR model, which has different algorithms than the GFS. Model refinement and development efforts would benefit from a focus on accurate representation of the diurnal cycle of temperatures as this basic characteristic of weather can reveal strengths and weakness in the model physics.

1 Introduction

Numerical weather prediction (NWP) models involve a suite of physical parameterizations, including convection, microphysics, land surface, boundary layer, and radiation schemes. The joint interactions among these parameterizations often yield difficulties in diagnosing sources of error within a model (e.g. Fovell et al., 2010; Bu et al., 2017; Caron & Steenburgh, 2020). The National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) and the High-Resolution Rapid Refresh (HRRR) models undergo a detailed testing and verification process before new operational versions are released (e.g. EMC Model Evaluation Group, 2019, 2020a, 2020b). NCEP's verification process focuses on aggregate statistics at the hemisphere, conterminous United States (CONUS), or CONUS sub-region scales and uses case studies to illustrate specific model strengths and weaknesses. For example, NCEP has documented a cold bias of approximately 0.5° C in CONUS East and 0.7° C in CONUS West within GFS v15 at approximately 36 hours that increases in magnitude with lead time (EMC Model Evaluation Group, 2020a).

We use a relational database to facilitate analyses for specific forecast and observed conditions. Examination of hourly model output across the diurnal cycle combined with conditioning on specific weather conditions provides a robust test of several aspects of model physics, and aids error diagnosis by constraining conditions when the errors occur.

Our verification methodology compares the model forecasts to *observations*, not to (re)analyses. A key weakness of reanalyses is that their accuracy is less well understood than the uncertainties in observations (Parker, 2016). Data assimilation methods often weigh observations less when they differ more from the model's solution (e.g. Houtekamer et al., 2005). Hence, uncertainties in reanalysis products are likely larger in locations and in weather conditions where numerical forecast models struggle—the very set of circum-

stances where information is most critical for model evaluation and refinement. The downside of using observations is that they are not available everywhere. By comparing the geographic distribution of errors, if a certain bias is present throughout much of the United States, it is more likely to be the result of a model physics weakness than an observation issue.

2 Data and Methods

Data from Automated Surface Observing System (ASOS) sites and GFS and HRRR models are compared for the period of November 1, 2019 to March 31, 2020. ASOS observations and matched model point data are stored in a relational MySQL database, which allows for easy querying of the data for analysis. Storing point data requires much less space than storing the full model gridded files. For this analysis, we use the following observed meteorological variables: 2-m temperature, 10-m wind speed, and sky condition. For each ASOS site, we obtain the corresponding GFS and HRRR model values of 2-m temperature, 10-m wind speed, and snow depth at each model run's set of valid times and lead times.

2.1 Observations

We used hourly Meteorological Terminal Air Reports (METAR) from 210 ASOS sites at airports in the CONUS to compare to model output. After data processing and quality control, variable values for each airport are uploaded to the database. The top-of-the-hour observations (i.e., no special observations) are compared to the model forecast valid at that hour. In situations where NOAA's Meteorological Assimilation Data Ingest System (MADIS) quality control was passed and our ingest determined the temperature value to be physically plausible, but the magnitude of the model temperature error was found to be greater than 20° C, that specific forecast hour and observation pair is not used in analysis. Sky conditions are delineated by ASOS as CLR $\leq$ 5%, 5% $<$ FEW $\leq$ 25%, 25% $<$ SCT $\leq$ 50%, 50% $<$ BKN $\leq$ 87%, and OVC $>$ 87% (NOAA, 1998).

The specific geographic coordinates chosen for each airport site were the approximate center of the airport property. This was a compromise between the ASOS site and the various discontinuity sensors used to make meteorological measurements at different points across the airfield (NOAA, 1998). Choosing a central location accounts for the unknown variation in exact locations used for measurements. For example, an airport may have multiple wind sensors and report only the value from the active runway. For 201 out of the 210 airport sites, we found the center of the airport property to be within 2 km of the ASOS station. For the other 9 airports, the ASOS site was within 3 km of the airport's geographic center.

2.2 Model Output

2.2.1 GFS

The operational versions of NOAA's GFS model model, which changed from version 15.1 to 15.2 on November 7, 2019 at 12 UTC (Maxson, 2019), were used for analysis. The absence of any major model changes with this update (Maxson, 2019) allows the entire date range to be analyzed in aggregate. All GFS model initialization times (0000, 0600, 1200, and 1800 UTC) were ingested into the database. We used the hourly GFS output for forecast hours 1 to 120. Since no long-term archive of the hourly output was known to exist, our own archive had to be created using the rolling 30-day archive on Amazon Web Services (AWS), which is part of NOAA's Big Data Program.

The 0.25 degree gridded GFS data were downloaded from the NOAA AWS Cloud. Spatial linear interpolation is used to obtain model values within the 0.25 degree grid

boxes at the 210 airport sites. The coarse resolution can yield mismatches between actual and modeled surface types for airports with runways adjacent to water. For example, the New York airports JFK and LGA are classified as water surface type rather than land.

2.2.2 HRRR

We use the High-Resolution Rapid Refresh (HRRR) version 3 to compare with the GFS model from November 1, 2019 to March 31, 2020. Only the 0000, 0600, 1200, and 1800 UTC initializations of the HRRR, which extend through forecast hour 36, are used to compare with the corresponding GFS initialization times. Since the effective grid length in this model is approximately 3 km (NOAA, 2020), the nearest grid point was chosen as being representative of the conditions at each of the airport sites. HRRR grids were downloaded from the University of Utah HRRR archive (Blaylock et al., 2017) and data at the nearest grid point to the 210 airports were used to populate the database for this study.

2.3 Diurnal Cycle

To address the diurnal cycle of temperature errors, we examine both hourly data and the temperature at the time of the winter climatological diurnal low and high temperatures at 7 a.m. and 3 p.m. local standard time (LT). We approximate 7 a.m. and 3 p.m. LT using longitude bands: Eastern time between 67.5° W and 82.5° W as 12 UTC and 20 UTC, Central time between 82.5° W and 97.5° W as 13 UTC and 21 UTC, Pacific time between 112.5° W and 127.5° W as 15 UTC and 23 UTC.

The local times of the climatological low and high temperature often do not coincide with the four times a day where exact 24-hour, 36-hour, 48-hour, etc. forecasts exist. The following analyses use the forecast lead times closest to, but less than, the desired lead time. Since we use forecasts initialized every six hours, 1 UTC corresponds to a 43 hour lead time, 2 UTC to 44 hours, 3 UTC to 45 hours, 4 UTC to 46 hours, 5 UTC to 47 hours and 6 UTC to 48 hours. This pattern is repeated every six hours. For brevity, we call all these a 48-hour lead time.

3 Results

We examine the daily average biases in surface temperatures within the GFS and HRRR over CONUS for the 210 airports in our relational database (Fig. 1). The GFS daily average cool temperature bias based on the airport locations of -0.70° C at 24-hours increasing in magnitude to -0.94° C at 120-hours closely matches the daily average biases found by NCEP (EMC Model Evaluation Group, 2020a). To help diagnose conditions when these biases are more frequent, we also examine biases at the times of the climatological low and high temperatures (7 a.m. and 3 p.m.). Figure 1 shows the clear diurnal variation in CONUS average temperature errors within GFS, with a much stronger cold bias during the daytime and a smaller cold bias at night. The HRRR has smaller temperature biases overall than GFS and has a cold bias during the night of approximately 0.5° C and a slight warm bias during the day.

We examined various weather conditions to determine in what circumstances stronger biases were more likely to occur. We found that for both GFS and HRRR, nighttime warm biases were usually larger in conditions with less cloudiness. Figure 2 shows the average model biases at 36-hour lead time for the subset of conditions when $\leq 50\%$ cloud cover is observed for each airport. For this subset of data, the CONUS average overnight temperature bias in the GFS is is 0.95° C as compared to 0.08° C for the HRRR model. When the data are further conditioned for low winds ($\leq$ 2.57 m/s or 5 kts), the CONUS average overnight low temperature error increases to 1.70° C for the GFS and increases

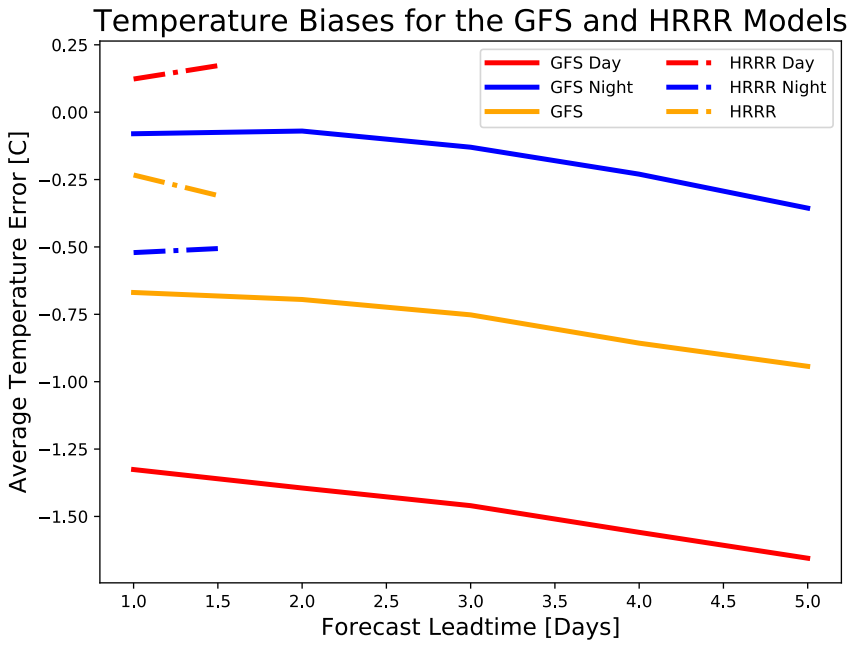

Figure 1. Results from analysis of both the GFS model (solid line) and the HRRR model (dash-dotted line) for various lead times at different times of day. Gold lines show the daily average bias. The daytime data (red) represents average CONUS temperature errors at the time of the diurnal high temperature (3 p.m. LT). The nighttime data (blue) represents the average CONUS error at the time of the diurnal low (7 a.m. LT). Data used are for all 210 airports in all weather conditions from 1 November 2019 to 31 March 2020. (Sample size: 152 days)

to 0.59° C for the HRRR (not shown). Conditions with low cloud cover and low winds are typically associated with strong nocturnal inversions.

The spatial patterns of errors are markedly different at the approximate time of the high temperature (3 p.m. local time) with $\leq 50\%$ cloud cover. There is a GFS cold bias at a 36-hour lead time in these conditions (Fig. 2a), with a CONUS average of -1.85° C. At the time of the daytime high, for periods with $\leq 50\%$ cloud cover, the HRRR tends to have a cold bias east of the Great Plains and a warm bias to the west. The HRRR also tends to have a slight cool bias overnight at coastal sites. Some of the largest cold biases in GFS and warm biases in HRRR are at airport locations in the intermountain west.

Overall, we identify a strong and obvious diurnal cycle of temperature errors under clear skies in the GFS model. The diurnal cycle of errors for the airports at Detroit, MI (DTW) and Oklahoma City, OK (OKC) are representative of many airports across the US (Fig. 3abcd). In the GFS, the temperature errors are smallest near sunset and increase overnight until the time of the climatological daily minimum temperature (Fig. 3ac). Once the sun comes up, the sign of the errors switches to negative (cool bias) during the day. For Oklahoma City, the GFS temperatures are approximately 3° C too high at night, and 1° C too low during the day. These patterns of temperature errors throughout the day are present at most sites across the United States in the GFS, and it is indicative of a substantial issue with the model in typical conditions of low sky cover. In contrast, the errors in the HRRR model do not yield much of a diurnal cycle in conditions with $\leq 50\%$ cloud cover (Fig. 3bd). Specifically, the median bias throughout the entire day in the HRRR model is close to 0° C for both Oklahoma City and Detroit.

The spatial pattern of errors overnight when observed cloud over is $\leq 50\%$ in the GFS indicates that some airports in the northern tier of the US have cold biases (Fig. 2). We examined the role of model snow cover in these errors by extracting the subset of data with observed cloud cover of $\leq 50\%$ and a model forecast of least a 1 cm snow depth (i.e. snow already on the ground). ASOS does not automatically record snow cover as it is typically augmented by a human observer at select airports (NOAA, 1998). Based on webcam footage, we found that if the model indicated snow depth > 1 cm, then snow cover was usually observed.

The diurnal cycle of temperature errors for MSP airport in Minneapolis, MN (MSP) for cloud cover $\leq 50\%$ and snow on the ground indicates a cold bias in both the GFS and the HRRR at all times of day. We see similar results for other northern cities with persistent snow cover such as FSD airport in Sioux Falls, SD and ABR airport in Aberdeen, SD. We speculate that some interactions between the snow covered surface and near-surface temperatures within the models are not being simulated properly.

4 Discussion and Summary

Accurate forecasts in conditions of low cloud cover have important practical applications. Errors of a few degrees high or low in forecasts for temperatures near 0° C are especially impactful for aviation and road transportation. Surface overnight low temperatures in winter are crucial for predicting whether frost will form and morning commute road conditions. Daytime high temperatures in conditions of low cloud cover are important for predicting to what extent existing snow and ice will melt. Planning for de-icing operations for roads and at airports benefits from 36-hour or more lead times.

By conditioning the NWP model error analysis on various cloud cover and wind speed conditions and times of day, we have revealed some major systematic issues in the GFS model temperature forecasts that are not as obvious in CONUS daily averages or in case studies. In particular, the GFS struggles to adequately represent the diurnal cycle of temperatures at 36-hour lead times under the mundane conditions of $\leq 50\%$ cloud

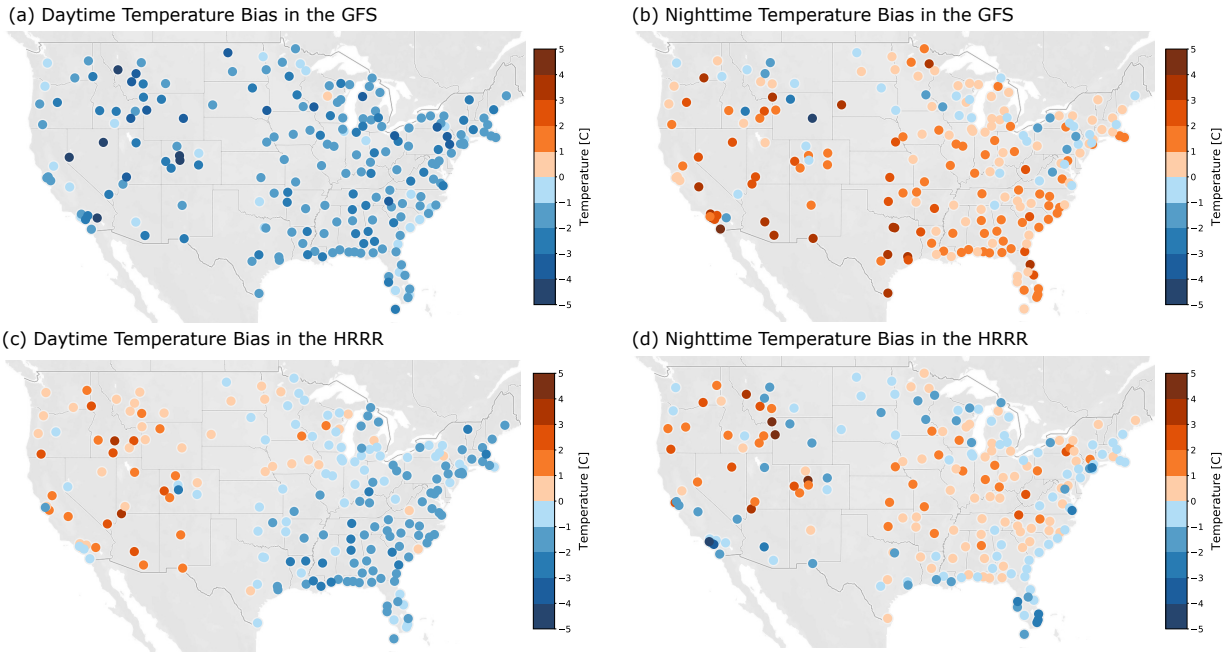

Figure 2. CONUS map of 210 airport sites showing the magnitude and sign of 36-hour lead time temperature biases (model - observation) with conditions of $\leq$ 50% cloud cover at the times of the winter climatological daily low temperature 7 a.m. LT and the daily high temperature 3 p.m. LT. a) GFS at 3 p.m. LT, b) GFS at 7 a.m. LT, c) HRRR at 7 a.m. LT, and d) HRRR at 3 p.m. LT. Red shading indicates the model is too warm, and blue shading indicates the model forecast is too cold in the November 1, 2019 to March 31, 2020 time frame. Sample size percentiles for both GFS and HRRR at 3 p.m. LT are 25%=49, 50%=65, 75%=84, and for both GFS and HRRR at 7 a.m. LT are 25%=50, 50%=63, 75%=77.

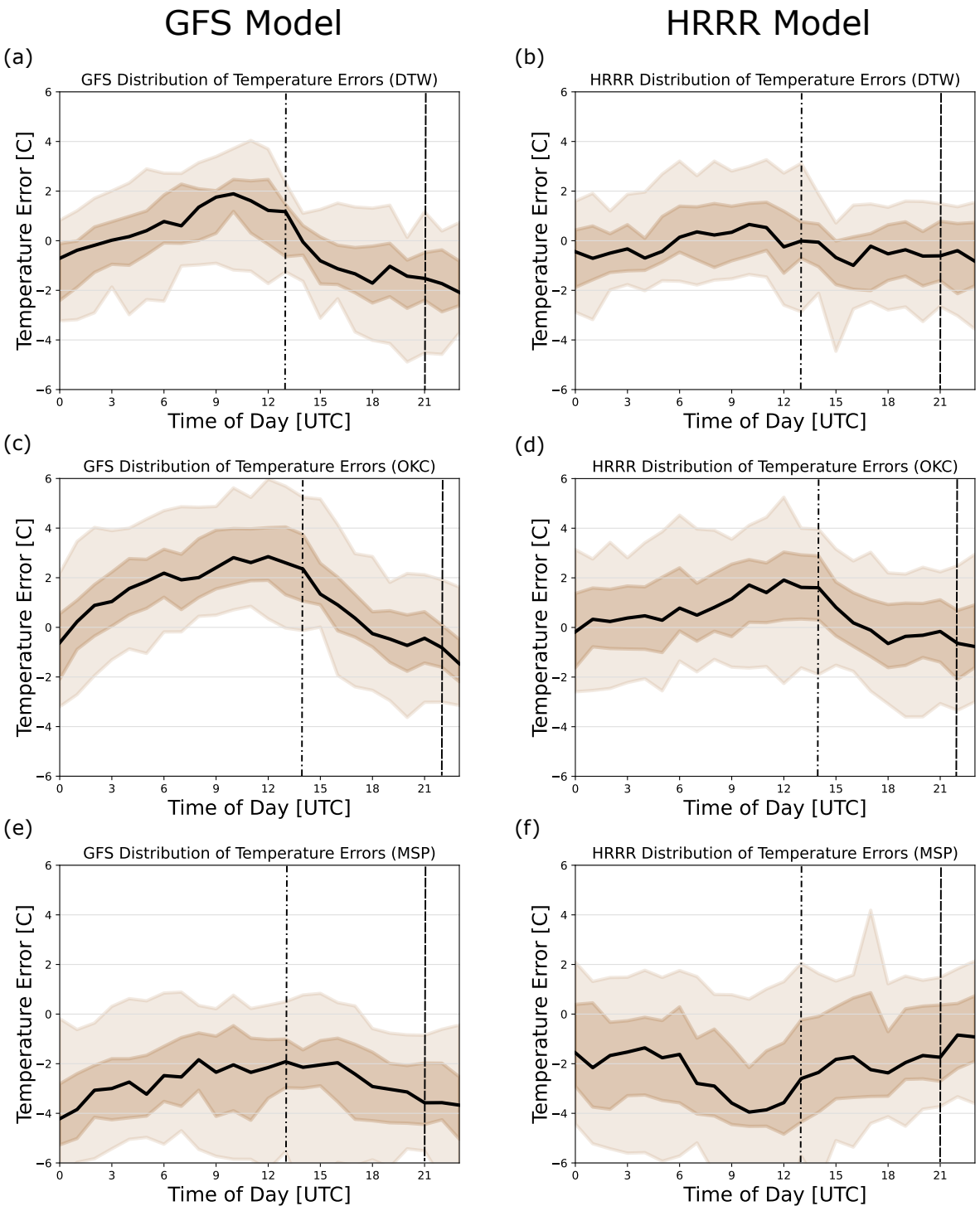

Figure 3. Distribution of temperature errors (model - observed) for conditions with observed $\leq 50\%$ cloud cover during the 36-hour forecast by time of day for (a,b) Detroit, MI, (c,d) Oklahoma City, OK, and (e,f) Minneapolis, MN. Errors in the GFS are shown by the left column (in a,c,e) and HRRR errors are shown in the right column (in b,d,f). The approximate time of the daily low temperature is indicated by the vertical dash-dotted line and the approximate time of the daily high temperature is indicated by a vertical dashed line. Dark beige shading extends from the 25th to 75th percentiles with the median indicated by the solid black line. Light beige shading spans the 5th to 95th percentiles. Minneapolis data (in e,f) further restricted to hours with forecast snow cover. Sample size distributions vary by location and time of day. Median sample sizes for OKC=65, DTW=29, MSP-HRRR=30, and MSP-GFS=36.

cover. Typically, the GFS is too warm overnight by 1° C and too cool during the day by almost 2° C. Overnight errors grow in magnitude when the $\leq 50\%$ cloud cover subset of data is further conditioned by low wind speeds for both the GFS and HRRR models. The strong diurnal cycle in errors need to be considered when making further model improvements. Otherwise, a parameterization refinement indicating only improvement in daily averages may yield degraded accuracy at different times of day.

The HRRR model does a better job at representing the diurnal cycle for $\leq 50\%$ cloud cover conditions and has smaller errors overall at 36-hour lead times. There are regional patterns in HRRR errors including a daytime-high cool bias in eastern US, a daytime warm bias in western US, and a nighttime cool biases at many coastal sites that merit further investigation.

Examination of errors by similar weather conditions on many days, rather than simple date ranges, helps constrain the portion of model physics in which larger forecast errors are more likely to occur. In this case, there appears to be an inadequacy in representing nocturnal temperature inversions in the interactions among radiation, boundary layer, and land surface parameterizations. To remedy this issue, portions of the suite of Unified Forecast System (UFS) medium-range parameterizations used in GFS v15 need to be improved before these physics packages are used in other models.

While the results presented here address only temperature errors, we demonstrate the utility of examining the diurnal cycle of errors and conditioning model to observation comparisons on weather characteristics to identify specific conditions when NWP model errors are larger. This type of model diagnosis is akin to identifying symptoms and aids in constraining where errors in parameterizations are more likely to reside. Resources for model refinement are limited. These types of analyses can aid in targeting investigations of error sources and revisions to physics packages to ensure models produce the right answer for the right reasons.

Acknowledgments

A rolling 30-day archive of GFS model output is available from NOAA via Amazon Web Services at https://s3.console.aws.amazon.com/s3/buckets/noaa-gfs-bdp-pds. HRRR model output is available through the University of Utah's HRRR archive (Blaylock et al., 2017). An archive of MADIS METAR data is available from NOAA at ftp://madis-data.ncep.noaa.gov/archive/. Data presented in figures is accessible at DOI 10.17605/OSF.IO/YTHR2. This work was supported in partnership with Delta Air Lines.

References

Blaylock, B. K., Horel, J. D., & Liston, S. T. (2017). Cloud archiving and data mining of high-resolution rapid refresh forecast model output. *Comput. Geosci.*, *109*, 43–50. doi: 10.1016/j.cageo.2017.08.005

Bu, Y. P., Fovell, R. G., & Corbosiero, K. L. (2017). The influences of boundary layer mixing and cloud-radiative forcing on tropical cyclone size. *J. Atmos. Sci.*, *74*, 1273-1292. doi: 10.1175/JAS-D-16-0231.1

Caron, M., & Steenburgh, W. J. (2020). Evaluation of recent NCEP operational model upgrades for cool-season precipitation forecasting over the western conterminous united states. *Wea. Forecasting*, *35*, 857-877. doi: 10.1175/WAF-D-19-0182.1

EMC Model Evaluation Group. (2019). *FV3GFS (GFSv15.1) Status Update: Addressing Excessive Snowfall and Low-Level Cold Bias.* (Accessed 27 May 2020, `http://www.emc.ncep.noaa.gov/users/Alicia.Bentley/fv3gfs/updates/MEG_4-04-19_FV3GFS_COLD.pptx`)

EMC Model Evaluation Group. (2020a). *GFSv16 MEG Evaluation Overview.* (Ac-

cessed 28 September 2020, https://www.emc.ncep.noaa.gov/users/meg/gfsv16/pptx/MEG_9-24-20_GFSv16_MEG_Eval_Overview.pptx)

EMC Model Evaluation Group. (2020b). *The MEG Perspective on the Implementation of RAPv5/HRRRv4.* (Accessed 4 February 2021, https://www.emc.ncep.noaa.gov/users/meg/rapv5_hrrrv4/updates/MEG_2020-12-03_RAPv5_HRRRv4_Implementation.pptx)

Fovell, R. G., Corbosiero, K. L., Seifert, A., & Liou, K.-N. (2010). Impact of cloud-radiative processes on hurricane track. *Geophys. Res. Lett.*, *37*, L07808. doi: 10.1029/2010GL042691

Houtekamer, P. L., Mitchell, H. L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., & Hansen, B. (2005). Atmospheric data assimilation with an ensemble kalman filter: Results with real observations. *Mon. Wea. Rev.*, *133*, 604–620. doi: 10.1175/MWR-2864.1

Maxson, B. (2019). *Upgrade NCEP global forecast systems to v15.2.0* (Service Change Notice No. 19-84). (https://www.weather.gov/media/notification/scn19-84gfs15_2.pdf)

NOAA. (1998). *Automated Surface Observing System (ASOS) user's guide* (User's Guide). (https://www.weather.gov/media/asos/aum-toc.pdf)

NOAA. (2020). *The High-Resolution Rapid Refresh (HRRR).* (Accessed 13 November 2020, https://rapidrefresh.noaa.gov/hrrr/)

Parker, W. S. (2016). Reanalyses and observations: What's the difference? *Bull. Amer. Meteor. Soc.*, *97*, 1565–1572. doi: 10.1175/BAMS-D-14-00226.1