Comparison between WRF-ARW and WRF-NMM objective forecast verification scores

L. Bernardet[1%&], J. Wolf[2%], L. Nance[2%], A. Loughe[1%♣], B. Weatherhead[1%♣], E. Gilleland[2%], B. Brown[2%]

[1]NOAA Earth System research Laboratory, Boulder, CO
[2]National Center for Atmospheric Research, Boulder, CO

I. Introduction

The Weather Research and Forecasting (WRF) model contains two dynamic cores: the Non-hydrostatic Mesoscale Model (NMM – Janjic 2003) core (developed at the National Centers for Environmental Prediction (NCEP) and the Advanced Research WRF (ARW – Skamarock et al. 2005) core, developed at the National Center for Atmospheric Research (NCAR). Each dynamic core corresponds to a set of dynamic solvers that operates on a particular grid projection, grid staggering, and vertical coordinate system. The WRF model also contains a multitude of physical parameterizations, many of which can be used with both dynamic cores.

This paper presents a comparison of temperature and precipitation forecast verification statistics for the ARW and NMM obtained as part of the Developmental Testbed Center (DTC, Bernardet et al. 2008a) 2007 13-km Core Test (Bernardet et al. 2008b). The main goal of this study is to determine if the inter-core differences increase with forecast lead time. This study is a follow up to the DTC 2006 Core Test (Brown et al. 2007), which compared ARW and NMM for 24-h forecasts and found no notable superiority in either core.

The ARW and NMM dynamic cores were used to forecast 120 cycles divided into the four seasons. The models were initialized every 36 h, resulting in alternating 00 and 12 UTC cycles. Details of the experimental configuration, results, and conclusions are presented in sections II, III, and IV, respectively.

II. Experiment Setup

The ARW and NMM 60-h forecasts were run on a CONUS domain with 13-km grid spacing and 58 vertical levels. The NMM used a 30-s timestep, while the ARW used a 72-s long timestep and a 18-s acoustic timestep.

Forecasts were computed for four seasons using data from: July and August 2005 for summer cycles, October and November 2005 for fall, January and February 2006 for winter, and March and April 2006 for spring.

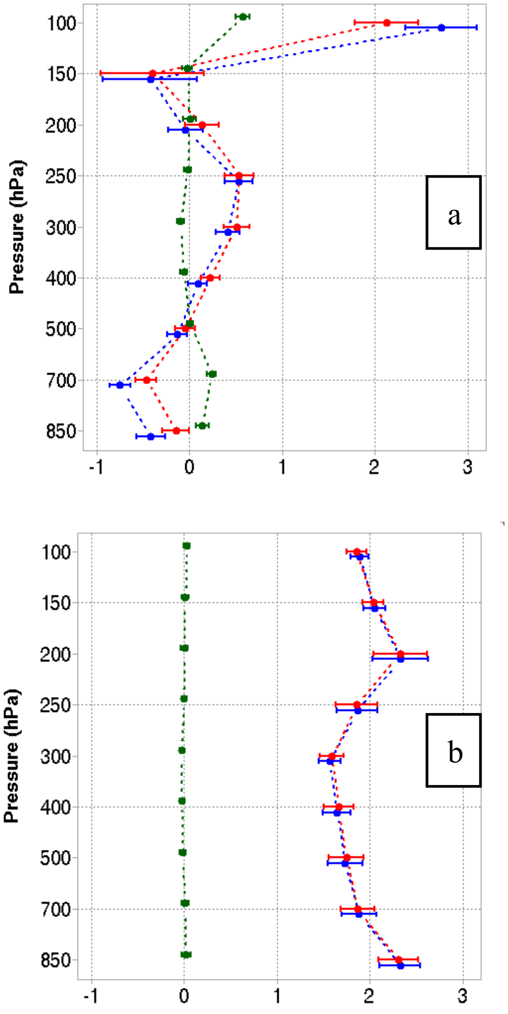

Figure 1. Vertical profiles of ARW (blue) and NMM (red) 60-h lead time temperature a) bias and b) BCRMSE (°C) with 99% CIs (horizontal bars). The ARW-NMM difference in absolute bias and BCRMSE are shown in green in a) and b), respectively.

[&] Contract with Systems Research Group, Colorado Springs, Colorado.

[%] Also affiliated with the Developmental Tested Center, Boulder, CO.

[♣] Contract with Cooperative Institute for Research in Environmental Sciences, Boulder, CO.

[*] Corresponding author address: Ligia Bernardet, NOAA Earth System Research Laboratory, Global Systems Division, 325 Broadway, Boulder, CO, R/GSD, email: ligia.bernardet@noaa.gov.

Both cores used the North American Mesoscale model (NAM, Eta model at the time) for cold-start initial and boundary conditions. The ARW and NMM were configured with an identical physics suite, including the following parameterizations: Ferrier microphysics, Janjic surface layer, Mellor-Yamada-Janjic planetary boundary layer, Betts-Miller-Janjic convection, Noah land-surface model, and GFDL radiation.

The WRF Postprocessing System (WPP - Chuang et al. 2004) was used to de-stagger the forecasts and to interpolate them to a common Lambert-Conformal grid. Additionally, the WPP was used to derive meteorological variables including mean sea level pressure, and to interpolate the forecasts to isobaric surfaces.

Using the NCEP Verification System (Chuang et al. 2004), forecasts were interpolated to the location of the observations (METARs and RAOBS) and used to generate partial sums averaged over the continental United States (CONUS). For the precipitation verification, a grid-to-grid comparison was performed against the River Forecast Center analyses valid at 12 UTC. From the results of the NCEP Verification System, several metrics were created and an aggregation in time over the entire Test period was computed. For brevity, the results presented in this abstract are limited to area bias and equitable threat score (ETS) for precipitation and bias and bias-corrected root mean square error (BCRMSE) for temperature. The BCRMSE represents the errors without bias and is defined as the square root of the estimated variance of the error which, when summed to the square of the bias, amounts to the mean square error.

The temporal aggregation used for precipitation was the mean. Confidence Intervals (CIs) on the mean were computed from standard error estimates using a correction for the autocorrelation. Confidence levels on the mean of temperature metrics are an estimate because the distributions are not exactly Gaussian due to, for example, the presence of some outliers. Temporal aggregation for precipitation was performed by accumulating results in a contingency table covering the entire period of the Test. Since the precipitation bias and equitable threat score distributions deviate significantly from normality, a bootstrap resampling method was applied to the data, and the adjusted percentile method was employed to obtain CIs (DiCiccio and Efron, 1996). Autocorrelation for the precipitation statistics was not an issue because the 00 and 12 UTC cycles are aggregated separately, leading to a 72-h separation between cycles. To determine the differences between forecasts, ARW minus NMM pair-wise differences of metrics for each forecast were computed, and temporal mean and CIs created with the parametric (temperature) and bootstrap (precipitation) methods.

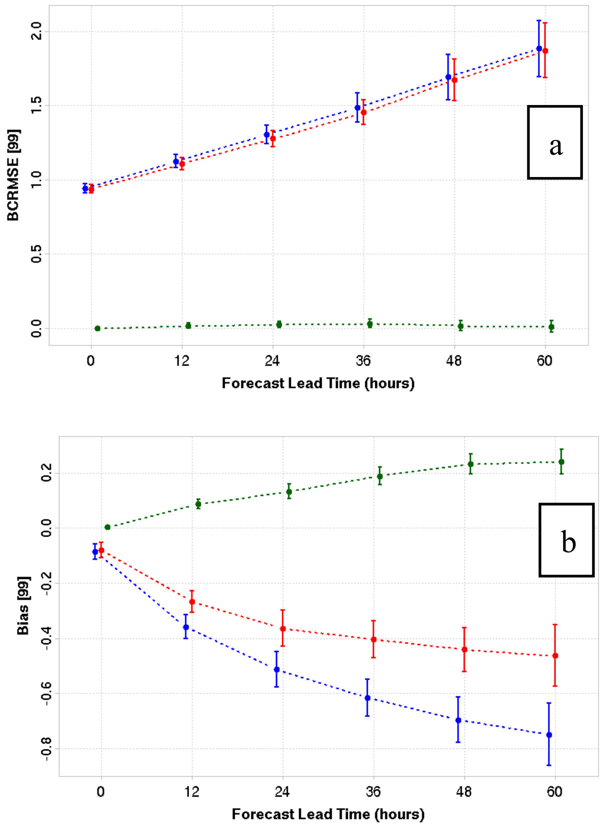

Figure 2. Time series of 700 hPa temperature a) bias and b) BCRMSE (°C) for ARW (blue) and NMM (red). The ARW-NMM differences in absolute bias and BCRMSE are shown in green in a) and b), respectively.

III. Results

A. Temperature

The vertical distribution of temperature bias for the average of the 12 and 00 UTC cycles for the 60-h lead time is shown in Fig. 1a. Both cores display cold forecasts at 850 and 700 hPa, topped by warm forecasts at 400, 300 and 250 hPa. A warm bias surpassing 2.0 °C is noted at 100 hPa. The inter-core differences in *absolute* bias, though small, reveal statistically significant differences at several levels. The ARW has better forecasts at 400 and 300 hPa, while the NMM has superior forecasts at 850, 700, and 100 hPa. The inter-core differences reach as much as 0.5 °C at 100 hPa, favoring the NMM.

Both cores have BCRMSEs that decrease with height from about 2.3 °C at 850 hPa to 1.5 °C at 300 hPa (Fig. 1b). Above this level, the errors increase up to 200 hPa, where a local maximum of 2.2 °C is noted. Inter-core differences in BCRMSE temperature do not exceed 0.1 °C and cannot be considered statistically significant.

The evolution of temperature verification statistics with forecast lead times for two selected levels (700 and 300 hPa) is shown in Figs. 2 and 3. At 700 hPa, a level for which both models have negative bias at all lead times, the bias for both cores becomes progressively more negative with time, but this progression is more accentuated for the ARW, leading to increasing NMM superiority with time (Fig. 2a). The temperature BCRMSE evolution (Fig. 2b), on the other hand, indicates that while the error grows in time for both models, their difference does not. At 300 hPa, both cores have positive bias at all lead times (Fig. 3a). The bias increases over the first 36 h of the forecast, but decreases thereafter. The ARW superiority is statistically significant at all lead times. The inter-core difference is constant in the first 24 h, but increases thereafter, as the ARW bias improves faster than the NMM's. The 300-hPa BCRMSE (Fig. 3b) results are similar to the 700-hPa ones, showing an increase of the errors with forecast lead times, but indicating no growth of the inter-core differences.

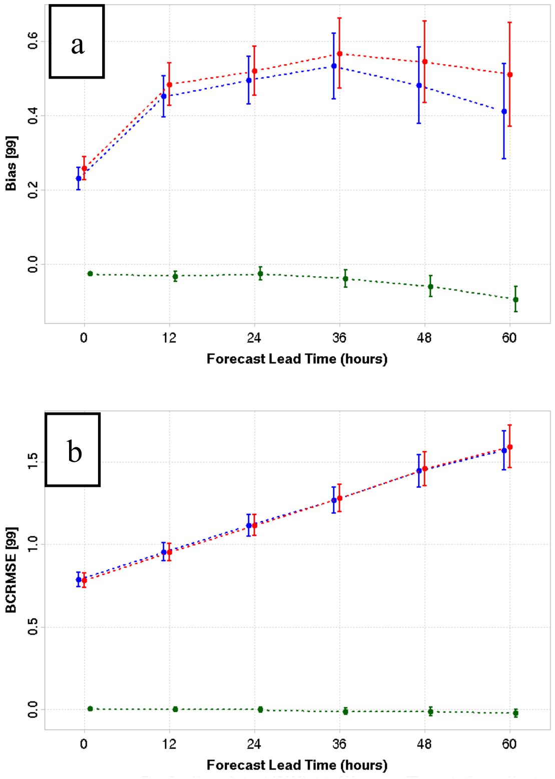

Figure 3. Same as Fig. 2, except using 300 hPa.

In contrast to the lower tropospheric results, when initialized at 00 UTC both cores have a surface warm bias at virtually all forecast lead times (Fig. 4a). The bias follows a prominent diurnal cycle, increasing during the night to reach a maximum in the early morning (15 UTC, equivalent to 9 AM CST). The bias then decreases during the day, to reach a minimum in the mid afternoon (21 UTC, equivalent to 3 PM CST). The positive bias increases from the first to the second night, but stabilizes by the third night. The inter-core difference in *absolute* bias shows a statistically significant difference at all but two forecast lead times, indicating an ARW superiority (smaller warm bias) of up to 0.5 °C. The results for the cycles initialized at 12 UTC (not shown) follow a similar pattern. The surface BCRMSE (Fig. 4b) shows an overall mild increase of the errors in time, with a semidiurnal modulation (errors increase in the early morning and early afternoon).

The inter-core BCRMSE differences follow a diurnal cycle but are very small and not statistically significant. No growth of the differences is observed towards the later forecast periods.

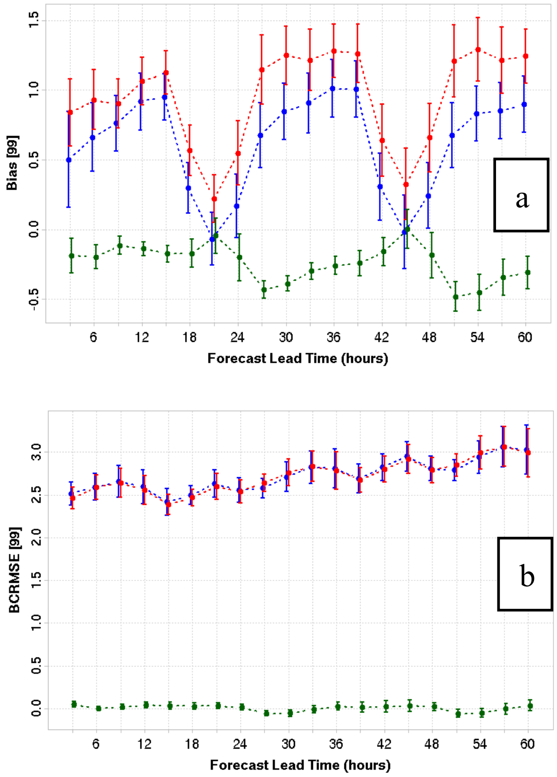

Figure 4. Same as Fig. 2, except with 2-m AGL.

B. Precipitation

The 24-h lead time bias and ETS for the 24h accumulated precipitation of the cycles initialized at 12 UTC are presented in Fig. 5 for several precipitation thresholds (from 0.01 to 2.0 in). For the 0.01 and 0.1 in thresholds, the ARW and NMM display overprediction. For the 0.5, 0.75, and 1.0 in thresholds, the NMM displays underprediction. No systematic error can be found for the other thresholds, as the CIs encompass 1. The ETS shows that the forecasts are best for the lower thresholds, and become progressively less accurate as the threshold increases.

The differences in bias and ETS for the 24-h forecast are also presented in Fig 5. Note that the CIs for the differences have been revised since Bernardet et al. (2008b). Statistically Significant (SS) bias differences can be noted for thresholds 0.01, 0.5, 0.75, 1.0 and 1.5 in. In all cases, the ARW produces a SS larger area of precipitation than the NMM. Given the values of bias, these differences represent a better NMM forecast for the 0.01 in threshold and a better ARW forecast for the 0.5, 0.75, and 1.0 in thresholds.

As the lead time progresses, the number of SS bias differences between the cores decreases. At the 36-h lead time, the bias difference is SS different only for thresholds 0.01, 0.75, and 1.0 in, at 48 h, only for 0.01 and 0.1 in and at 60 h, only for 0.01 in (Fig. 6). All differences but one occur because the ARW has more areal coverage. Of all differences at lead times 36-, 48-, and 60-h, two favor the NMM and one, the ARW.

The only SS difference in ETS appears for the 0.01 in threshold at the 36-h lead time, favoring the ARW (not shown).

IV. Conclusions

Temperature and precipitation forecast verification results from the DTC 2007 13-km Core Test were presented. Results from 24-h accumulated precipitation indicate that SS differences bias exist, with all but one case indicating the ARW has larger precipitation coverage. Since both cores tend to underpredict at lower thresholds and overpredict at some higher thresholds, this results in an equal number of ARW and NMM superior forecasts. The bias differences decrease with forecast lead time. There are virtually no SS ETS differences.

The temperature BCRMSE results also indicate virtually no difference between the cores, and no growth of the difference in time. The temperature bias results are more complex.

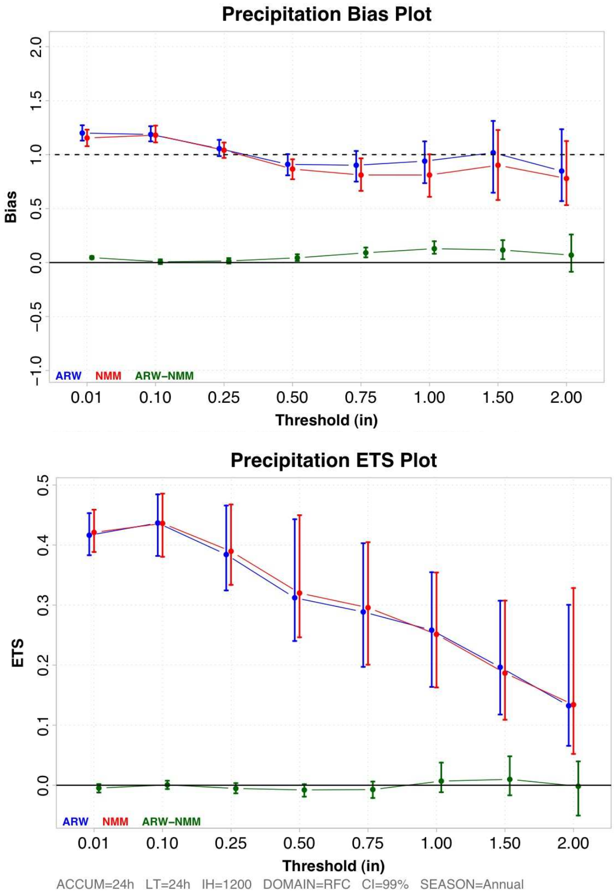

Figure 5. a) Precipitation bias and b) ETS with 99% CIs for the 24-h lead time for ARW (blue) and NMM (red). The ARW-NMM pairwise difference is shown in green.

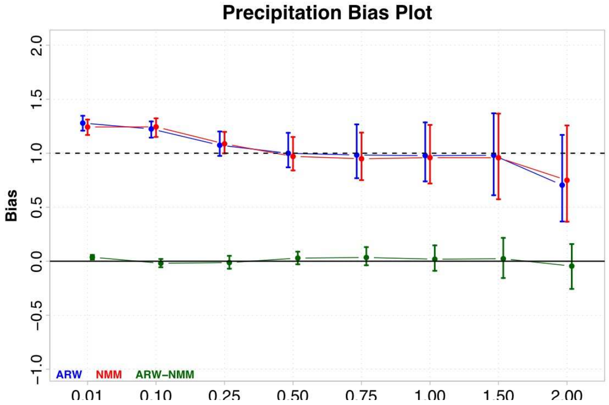

Figure 6. Precipitation bias with 99% CIs for the 24-h lead time for ARW (blue) and NMM (red). The ARW-NMM pairwise difference is shown in green.

Small, but statistically significant, absolute bias differences between the cores, some favoring ARW and others favoring NMM, were presented for several vertical levels at the 60-h lead time. In the time series for the levels presented (2-m AGL, 700 hPa and 300 hPa), the sign of the difference remains virtually constant throughout the forecast. At 300 hPa, the inter-core difference does not grow in time but at 2-m and 700 hPa growth is observed.

To reach a final conclusion about the existence of significant differences between the ARW and NMM forecasts out to the 60-h lead time, future work will include an examination of other variables, such as humidity, and winds. Additionally, the results will be stratified by regions of the CONUS and by season. Additionally, an analysis based on the median bias and BCRMSE will be done to complement the mean bias and BCRMSE presented here, with the goal of reaching a more robust result that is not sensitive to a few outliers.

Acknowledgements: The authors would like to thank the following scientists who participated in the Core Test: Barbara Brown, Randy Bullock, Robert Gall, John Halley Gotway, Lacey Holland, Beth Weekley, and Lara Ziady of NCAR, and Steven Koch and Jeff Smith of NOAA ESRL. Additionally, the authors would like to thank the following WRF developers for their assistance in configuring the model and/or developing code for this test: Michael Duda, Jimy Dudhia, Dave Gill, and Joe Klemp from NCAR, Tom Black, Hui-ya Chuang, Geoff DiMego, Michael Ek, Zavisa Janjic, and Matthew Pyle from NOAA NCEP.

V. References

Bernardet, L., L. Nance, M. Demirtas, S. Koch, E. Szoke, T. Fowler, A. Loughe, J. L. Mahoney, H.-Y. Chuang, M. Pyle, and R. Gall, 2008a: The Developmental Testbed Center and its Winter Forecasting Experiment. *Bull. Amer. Meteor. Soc.,* **89**, 611-627.

Bernardet, L., J. Wolff, L. Nance, E. Gilleland, B. Weatherhead, C. Harrop, M. Govett and A. Loughe, 2008b. Objective verification results from forecasts generated with the ARW and NMM dynamic cores of WRF. 8th WRF Users' Workshop, Boulder, CO, June 23-27.

Brown, J. M., S. Benjamin, T. G. Smirnova, G. A. Grell, L. R. Bernardet, L. B. Nance, R. S. Collander, and C. W. Harrop, 2007: Rapid-Refresh Core Test: aspects of WRF-NMM and WRF-ARW forecast performance relevant to the Rapid-Refresh application. *22st Conference on Wea. Anal. Forec.*, Park City, UT, Amer. Meteor.Soc.

DiCiccio, T.J. and Efron, B., 1996. Bootstrap confidence intervals, Statistical Science, 11(3):189--228.

Chuang, H., G. DiMego, M. Baldwin, and WRF DTC team, 2004. The NCEP WRF post-processing software. *The Joint WRF/MM5 Users Workshop*, Boulder, CO.

Janjic, Z. I., 2003. A nonhydrostatic model based on a new approach. *Meteorol. Atmos. Phys.* **82**, 271-–285.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang and J. G. Powers, 2005: A Description of the Advanced Research WRF Version 2, NCAR Tech Note, NCAR/TN–468+STR, 88 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO, 80307]. Available on-line at http://box.mmm.ucar.edu/wrf/users/docs/arw_v2.pdf).