# Summaries of Statistical Analyses of Differences in Relative Humidity, Temperature, Wind and Precipitation between ARW and NMM WRF Cores using both Sonde and Aircraft Data.

Elizabeth Weatherhead, Gregory Noonan, Tressa Fowler, Ligia Bernardet, Louisa Nance and Steve Koch

This report summarizes detailed statistical analyses from three reports. These three reports are available at http://www-frd.fsl.noaa.gov/mab/wrfrr/doc or at http://www.dtcenter.org/ , or they may be obtained directly from the lead author, Betsy Weatherhead.

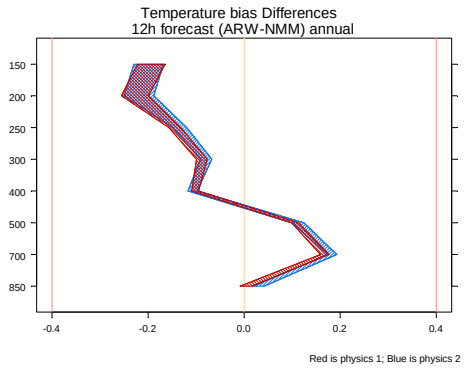
Weather forecasts were produced simultaneously using two different cores (ARW and NMM). The primary question addressed is whether one core is better than the other. For most parameters, there are small differences between the output of these two models. Two separate questions may be asked: are these differences large enough to be important and are these differences statistically significant. This report addresses the second question. We focus on large regions and examine the differences in temperature, relative humidity, winds and precipitation using advanced statistical techniques. The basic conclusions are presented in this report with considerable supporting details in the appendices.

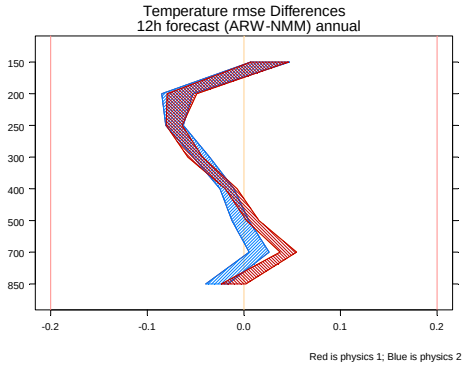
The models were, for the most part, run on the same days, using similar input. In general, when one model had a high bias for a particular forecast, the other model did as well. The design of the intercomparison as well as the common variance observed allows for a pairwise comparison of results from the ARW and NMM cores. The pairwise comparison is a statistical test that can be used when there is common variance between datasets. In this case, when the ARW bias was high, the NMM bias was also typically high, although not necessarily identically high. The common variance can be removed making the pairwise test more sensitive to high difference between the ARW and NMM forecasts. The pairwise comparison tests were calculated for three regions: CONUS, CONUS West and CONUS East. The tests were additionally carried out for 12 hour forecasts as well as 24 hour forecasts. All cases were examined for both Phase-1 and Phase-2 physics packages. Results are presented in both tabular and graphical form. All of these values are calculated using the full set of available data as well as the data divided by season: Winter, Spring, Summer and Autumn.

The pairwise comparison fundamentally looks at differences of pairs of numbers. In its most basic form, the pairwise test assumes that these differences are normally distributed, independent and of equal accuracy. The actual differenced data were examined for these three properties. The impact of deviation from these assumptions is discussed below. Based on these comparisons, we come to the following conclusions:

*Temperature Bias*

For both 12h and 24h forecasts, the lowest level examined, 850mb, both ARW and NMM have similar magnitude of bias. For 700mb and 500mb, the NMM bias is statistically significantly smaller than ARW. For 400mb and above, the ARW magnitude of bias is statistically significantly smaller than NMM. More information on this subject is available in Appendix 3 of Report 1.




*Temperature RMSE*

For both 12h and 24h forecasts, the ARW and NMM forecasts show a slightly smaller RMSE for the lowest level of the atmosphere examined, 500mb. This difference is small, but significant for CONUS West and near zero for CONUS East. For 300mb, 250mb and 200mb, the ARW shows a statistically significant smaller RMSE. In Summer, there is very little difference between the two forecasts. These results are independent of which physics package is used. More information on this subject is available in Appendix 4 of Report 1.

*Relative Humidity Bias*

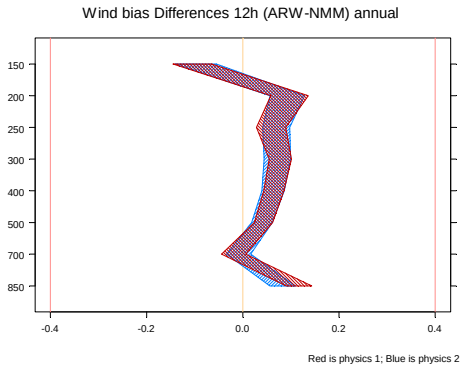
For both the 12h and 24h forecasts, the NMM has a smaller magnitude of bias than the ARW at 500mb. At 700mb the difference in bias is highly dependent on the physics package. The smallest difference between the two cores is observed at 850mb. Seasonally, the largest scatter is observed in the Summer. Most of these differences are highly statistically significant. More information on this subject is available in Appendix 5 of Report 1.

*Relative Humidity RMSE*

For both 12h and 24h forecasts, there is little difference between the RMSE for the ARW and NMM forecasts. This is particularly true at 500mb. At 850mb, ARW shows a statistically significant smaller RMSE for CONUS West. This is most true for Phase-2 physics. These differences are largest for Autumn and Winter. More information on this subject is available in Appendix 6 of Report 1.

*Wind Bias*

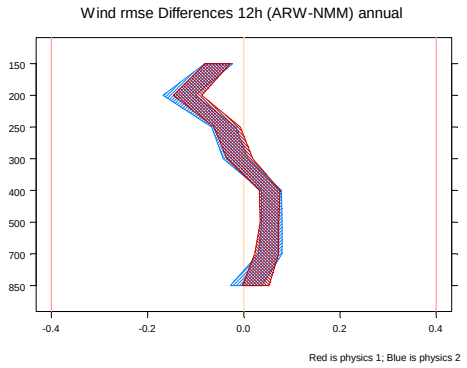
Wind bias is in general consistent from 850mb through 200mb although NMM shows a smaller magnitude of bias for these levels. At 150mb, ARW shows a smaller magnitude of bias. These results are somewhat seasonally dependent, particularly for the behavior above 300mb. For Winter and Autumn, ARW shows a much smaller magnitude of bias than NMM. More information on this subject is available in Appendix 7 of Report



1.

*Wind RMSE*

For both the 12h and 24h forecasts, NMM exhibits a smaller RMSE for the lowest part of the atmosphere (between 850mb and 300mb). Above 300mb, ARW has a smaller RMSE for all regions and all physics packages. This difference is most clear for the 12h forecast and becomes less clear with the 24h forecast, although both forecasts show similar patterns. These patterns are repeated seasonally, although at 850mb in Autumn and Winter, ARW has a smaller RMSE. More information on this subject is available in Appendix 8 of Report 1.




A different approach was used to assess core differences with precipitation forecasts. An advanced bootstrap method was used, giving the following results:

*Precipitation Bias*

For low precipitations thresholds, both models have bias greater than one, with the ARW models showing a larger bias. For thresholds greater than 0.3, both models using both physics packages, have bias values less than one. For these higher thresholds, ARW has a larger bias than NMM. For Phase-1 physics, ARW has a higher bias than NMM for all threshold levels. For Phase-2, ARW has a higher bias for threshold levels below 0.4. While the differences for Phase-1 are very small, the differences between ARW and NMM for both physics packages are statistically significant at the 95% confidence level. More information on this subject is available in Appendix 9 of Report 1.

*Precipitation Equitable Threat Score*

For Phase-1 physics, NMM has a higher equitable threat score for very light precipitation thresholds. For higher thresholds within Phase-1, ARW has a higher equitable threat score. For Phase-2 physics at all threshold levels, ARW has a higher equitable threat score. This difference is largest and strongest in statistical significance for light precipitation and is not statistically significant at the 95% level for high precipitation threshold levels. More information on this subject is available in Appendix 10 of Report 1.

*Model Initialization.*

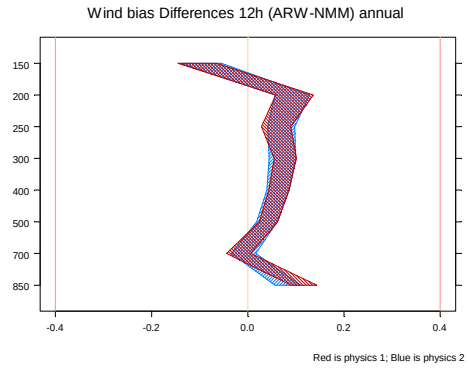
The pairwise approach was also used to examine whether the ARW and NMM models were initialized similarly. Based on an examination of temperature, relative humidity and winds it was clear that the initialization of the models was statistically different. While the differences were smaller than the differences observed with 12 and 24 hour forecasts, they were both statistically significant and showed systematic patterns with variables and height. More information on this subject is available in Appendix 11 of Report 1.

*Diurnal Differences in Forecasts.*
To assess whether the models performed differently throughout the day, the 12 hour and 24 hour forecasts were examined separately for 00Z and 12Z forecasting run times. If there was a very strong diurnal bias in either the ARW or the NMM model, we would expect to see significant differences between the difference in bias for the forecasts made for 12Z (e.g. the 24 hour forecast made at 12Z, or the 12 hour forecast made at 00Z) and the forecasts made for 24Z (e.g. the 24 hour forecast made at 00Z, or the 12 hour forecast made at 12Z). These differences were not detected in this comparison. The 12 hour forecasts show more similarity with each other, independent of what time of day was being forecasted. More focused studies to look at diurnal issues may uncover details not available from this approach. More information on this subject is available in Appendix 12 of Report 1.

The summary of the difference in model results has been calculated by region (CONUS, CONUS West and CONUS East), by season, by altitude, and for two separate physics packages. However, the mean differences average over many unique meteorological conditions over large areas. One attempt to understand how the results may vary for individual days is attempted by looking at the dependence of the differences in bias and RMSE. These results indicate that the difference in bias and RMSE are, in some cases, dependent on the meteorological conditions. More information on this subject is available in Appendix 15 of Report 1.

To assess the dependence of the wind results on sonde measurements, analyses were also performed using the wind results from Aircraft data. The results are presented in Report 2 and are summarized here:

*Wind Bias*

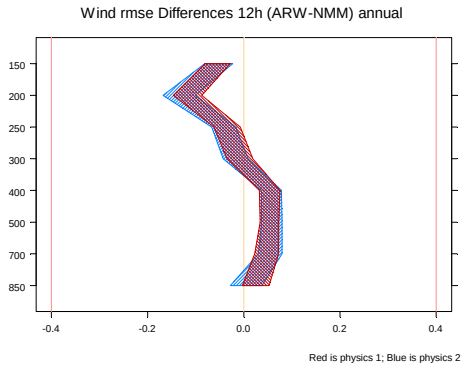



Wind bias results from sondes are in general consistent from 850mb through 200mb although NMM shows a smaller magnitude of bias for these levels. At 150mb, ARW shows a smaller magnitude of bias. These results are somewhat seasonally dependent, particularly for the behavior above 300mb. Aircraft data generally show that NMM has a smaller magnitude of bias, at all levels, although the difference in bias is smaller. The major difference is in the highest altitude. Aircraft results were analyzed up to 200-250mb. There are no aircraft results which correspond to the highest sonde level (150mb) where sondes show that ARW has a smaller magnitude of bias. In Autumn and Winter, measurements from ARW - NMM show different results for 200mb, with the sondes showing that ARW has a much lower magnitude of bias than NMM while the aircraft show a small advantage to NMM. These differences in bias are based exclusively on differences in forecasts for the locations at which observations are

available and not on the value of the measurements themselves. More information is on this subject is available in Appendix 3 and Appendix 4 of Report 2

*Wind RMSE*

For both the 12h and 24h forecasts, NMM exhibits a smaller RMSE for the lowest part of the atmosphere (between 850mb and 300mb). Above 300mb, ARW has a smaller RMSE for all regions and all physics packages. This difference is most clear for the 12h forecast and becomes less clear with the 24h forecast, although both forecasts show similar patterns. These patterns are repeated seasonally, although at 850mb in Autumn and Winter, ARW has a smaller RMSE. The results are similar from both sondes and aircraft, although the aircraft do not offer results for levels as high as 150mb. The seasonal information is the same for both aircraft and sondes, although the error bars from the aircraft data are smaller, presumably because the results are based on more observations. For 24 hour forecasts, the RMSE values are smaller for ARW than NMM for almost all levels from the aircraft data, while this is less true for the sonde data. This difference is particularly notable in Autumn and Winter. More information is on this subject is available in Appendix 5 and Appendix 6 in report 2.




*Model Initialization.*

The pairwise approach was also used to examine whether the ARW and NMM models were initialized similarly. Based on an examination of sonde winds it was clear that the initialization of the models was statistically different. However, the RMSE differences between ARW and NMM at initialization are extremely small when the aircraft data are used. Using sonde data, many of the differences while small, were highly statistically significant. Using aircraft data, we observe even smaller differences with most being statistically significant. More information on the model initialization of winds is available in Appendix 7 in report 2.

*Dependence of Results.*

An important question, when sonde data and aircraft data give different results, is why the results are different. In order to address this question we look at inherent biases in each of the datasets. For sonde data, it has been proposed that lost sondes at high altitudes could affect the results. We examine this in a stepwise approach. First it is verified that fewer measurements are available at 150mb than at 200mb. On average there are a few less sondes for the higher altitudes, but this difference is very small. Second, it is shown that at high altitudes (150mb and 200mb), high average zonal winds are associated with lower numbers of sonde measurements; this is not true lower in the atmosphere where the number of sonde measurements is roughly independent of average winds. Finally, we note that the estimates of RMSE and the differences in RMSE between ARW and NMM are roughly independent of the number of measurements collected. Thus, while the dropoff rate with sondes is real, and likely due to the ambient

conditions, it does not appear to have a large influence on the conclusion regarding RMSE and bias. More information is available on this in Appendix 8 in Report 2.

To assess the dependence of the temperature results on sonde measurements, analyses were also performed using the temperature results from Aircraft data. The results are presented in Report 3 and are summarized here:

*Temperature Bias*
Temperature bias results from aircraft are show that at the lowest level: below 850 mb, the ARW has a slightly smaller bias than NMM, with this difference being greatest for CONUS West. This is true for all four seasons, both physics packages and for both 12 and 24 hour forecasts. From 850mb through roughly 300 mb, NMM shows a smaller bias—again consistent for physics packages, regions of the country and for both 12 and 24 hour forecasts. The notable exception is in the summer when the results are quite mixed across regions. Examination of the raw bias numbers, shows that there is a general consistency of bias with season and height. Summer bias values for 24 hour forecasts, with both phase 1 and phase 2 physics are very small, resulting in the observed mixed results for differences in bias.

*Temperature RMSE*
In summer, temperature rmse differences from aircraft show that differences are near zero with NMM showing the smaller RMSE, particularly for CONUS West. For the other three seasons, NMM has a smaller RMSE around 500mb while ARW has a smaller RMSE above 300mb. These differences are generally smaller than 0.05 degrees and only sometimes significant at the two sigma level.

*Model Initialization.*
The pairwise approach was also used to examine whether the ARW and NMM models were initialized similarly. Based on an examination of sonde winds it was clear that the initialization of the models was statistically different. However, the RMSE differences between ARW and NMM at initialization are extremely small when the aircraft data are used. Using sonde data, many of the differences while small, were highly statistically significant. Using aircraft data, we observe even smaller differences with most being statistically significant. More information on the model initialization of winds is available in Appendix 7 of Report 3.

Further discussion of the statistical approach used and how it compares to other methods is presented in Appendix 13 of Report 1. A discussion of some of the assumptions associated with the statistical approach, and how they were tested and accounted for is presented in Appendix 14 of Report 1.

**References**

Von Storch, H. and F. Zwiers, “Statistical Analysis in Climate Research”, Cambridge University Press, 1999.

Rowlings, J., S Pantula and D. Dickey, “Applied Regression Analysis,” Springer-Verlag, 1998.

Neter, J., W. Wasserman and M. Kutner, “Applied Linear Statistical Models,” Irwin, 1990.

Wilks, Daniel S., “Satistical Methods in the Atmospheric Sciences,” Academic Press, 1995.

Box, G., G. Jenkins, and G. Reinsel, “Time Series Analysis: forecasting and control,” Prentice-Hall, 1994.

Weatherhead, E., et al., “Statistical considerations in the detection of trends.” JGR, 103, 17,149-17161.

Efron, B. and R. Tibshirani, “An Introduction to the Bootstrap,” Chapman and Hall/CRC, New York, 1994.