Army Research Laboratory

Performance Comparison of High Resolution Weather Research and Forecasting Model Output with North American Mesoscale Model Initialization Grid Forecasts

by John Raby, Jeff Passner, Gail Vaucher, and Yasmina Raby

ARL-TR-6000 **May 2012**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer’s or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
White Sands Missile Range, NM 88002-5501

ARL-TR-6000 **May 2012**

Performance Comparison of High Resolution Weather Research and Forecasting Model Output with North American Mesoscale Model Initialization Grid Forecasts

John Raby, Jeff Passner, Gail Vaucher, and Yasmina Raby
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* May 2012	**2. REPORT TYPE** Final	**3. DATES COVERED (From - To)** October 2010−September 2011

4. TITLE AND SUBTITLE Performance Comparison of High Resolution Weather Research and Forecasting Model Output with North American Mesoscale Model Initialization Grid Forecasts	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) John Raby, Jeff Passner, Gail Vaucher, and Yasmina Raby	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Laboratory Battlefield Environment Division Computational and Information Sciences Directorate, ARL (RDRL-CIE-M) White Sands Missile Range, NM 88002-5501	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TR-6000
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Model Assessment Project used the National Center for Atmospheric Research (NCAR) Model Evaluation Tools (MET) software to compute traditional error statistics which compared the performance of the high resolution Weather Research and Forecasting (WRF) model with the North American Mesoscale (NAM) model. Gridded forecasts from these models were compared to weather observations from Meteorological Assimilation Data Ingest System (MADIS) mesonet observational data and National Centers for Environmental Prediction (NCEP) PrepBUFR observational data. MET Point-Stat was used to generate statistics based on the point differences between model forecasts and the observations. The statistics were aggregated to produce overall summary statistics for nine surface and five upper air meteorological variables to characterize the performance over the 31-day study period. The WRF forecasts were generated by the Nowcast Modeling Project in support of a common research objective to compare the performance of the 1-km and 3-km resolution WRF to that of the model used to initialize the WRF, which is the 12-km resolution NAM model. These model runs were executed over two domains centered over Dugway Proving Ground, Utah. The statistical results are discussed with some tabular and graphical examples and conclusions are offered which describe the results of the comparison.

15. SUBJECT TERMS

WRF, weather, forecast, validation, assessment, model, statistics, operations, evaluation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					John Raby
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)*
Unclassified	Unclassified	Unclassified	UU	72	(575) 678-2004

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

Acknowledgments

The error statistics included in this data report were generated using the Model Evaluation Tools (MET) developed at the National Center for Atmospheric Research (NCAR) through a grant from the United States Air Force Weather Agency (AFWA).

In addition, AFWA provided funding and guidance to the U.S. Army Research Laboratory (ARL) for Numerical Weather Prediction (NWP) model research.

Special appreciation is extended to Mr. Robert Dumais of ARL/Battlefield Environment Division (BED) who provided his interpretation and modeling insight and the source of the archived North American Mesoscale (NAM) forecast data.

Special appreciation is extended to Mr. Robert Flanigan of ARL/BED who provided system administration expertise which enabled the automation of the model evaluation processes necessary to accomplish the calculation of error statistics using the MET software.

Executive Summary

Problem

Weather has a significant impact on Army personnel, weapons, tactics, and operations; so accurate weather forecasts can be a deciding factor in any conflict, large or small. The weather forecasting task has shifted from a human forecaster located in-theater to computerized Numerical Weather Prediction (NWP) with the human forecaster located far from the area of interest.

Weather forecast validation has always been of interest to the civilian and military weather forecasting community. This interest has recently shifted from the accuracy of human forecasters to the accuracy of the NWP models. The validation of the models, especially high resolution models produced by NWP, has proven to be especially difficult when addressing small time and space scales.

The U.S. Army Research Laboratory's (ARL)'s Battlefield Environment Division (BED) utilizes a high resolution model, the Weather Research and Forecasting (WRF) model, that will use the Four Dimensional Data Assimilation (FDDA) technique in future situations. This system will ingest battlefield weather observations to improve the quality of the forecasts. In order to show the value added of these forecasts over those produced by the standard WRF initialization forecast grid, it is necessary to first quantify the value added of the WRF model, which serves as the engine for the WRF-FDDA. Assessing the accuracy of this WRF model will provide a benchmark that will serve as the basis to determine, by comparison, if the forecasts from WRF-FDDA model are more accurate.

Results

ARL has performed case studies investigating the performance of various NWP models to develop appropriate weather forecast applications predominantly for military use. Previous studies have included some traditional statistical measures including Bias or Mean Error (ME), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) values. Resource constraints have dictated that only a small number of data points could be used for the calculations. A cornerstone study incorporated the use of the Model Evaluation Tools (MET) for the first time at ARL. This study enabled more comprehensive statistical evaluations of current Department of Defense (DoD) options for using the WRF model using a forecast grid-to-point observation model verification method (Sauter, 2009).[1]

[1] Sauter et al. *Traditional Statistical Measures Comparing Weather Research and Forecast Model Output to Observations Centered Over Utah;* ARL internal report; U.S. Army Research Laboratory: White Sands Missile Range, NM, 2009.

In a follow-on assessment, those previously developed methods were used as the core for a more robust implementation of MET in 2010, which included a large increase in Meteorological Assimilation Data Ingest System (MADIS) observational data. The MET process was also automated to generate ME, MAE, and RMSE statistics based on 140 WRF runs, which used seven different parameterization settings and two different WRF resolutions for 20 case study days between March 2009 and July 2010. This enhanced set of verification statistics, using grid-to-point evaluations, revealed that the errors varied, depending on the case study day. This case study approach may have skewed the results towards model performance in these specific types of weather conditions. Another conclusion drawn from this extensive set of statistics was that the different parameterization settings had little impact on the value of the error statistics with the possible exception of one of the boundary layer parameter settings. All settings essentially produced the same error values. In addition, it was found that the errors for the 1-km and 3-km horizontal resolution WRF versions were statistically the same (Raby, 2011).[2]

To address the concerns generated by the previous assessment about possible Biases introduced by the weather-driven case study approach, the WRF model was run at 3-km resolution and 1-km resolution during a continuous 31-day period from 27 May to 26 June 2011. To establish a benchmark comparison for WRF performance, which could later be compared with the WRF-FDDA performance, it was decided to acquire and assess the output of the North American Mesoscale (NAM) model. The NAM is a variant of the WRF model. The NAM model serves as the initialization grid for the WRF, but at a lower horizontal resolution of 12 km. Assessing the errors of the NAM model would then provide a basis for comparison of the WRF errors, to show if there was a value added in running the two higher resolution WRF models. This way, when the assessment of the WRF FDDA is conducted and a similar value added quantified, there would be a basis to determine which provides the larger value added over the same initialization grid.

Unlike the previous assessment, which compared error statistics aggregated over a 24-h period for 20 distinct case study days, the tact taken for this assessment was to aggregate the statistics generated using the grid-to-point method over all 31 consecutive model runs, to produce overall errors for comparison. For surface meteorological variables, these comparisons were accomplished using two different aggregations. The first was to aggregate the errors over all the runs and output single statistics which characterize the model errors for all 24 h of the model run over the 31-day study period. For these statistics, the 95% bootstrap confidence intervals were extracted to capitalize on the extensive sample size in order to obtain a reasonable assessment of how the three models compare and if the observed differences were statistically significant. The second was to aggregate the errors over all the runs and output the statistics by forecast hour. For upper-air meteorological variables, the model comparisons were accomplished by aggregating the errors generated from the differences between data collected from two

[2] Raby et al. *Traditional Statistical Measures Comparing Weather Research and Forecast Model Output to Observations Centered Over Utah*; ARL-TR-5422; U.S. Army Research Laboratory: White Sands Missile Range, NM, 2011.

rawinsonde upper-air observation sites and the model forecasts at various levels in the atmosphere.

A longer-term goal is to use the fuzzy and spatial verification capabilities in MET, to assess model performance not captured by traditional, grid-to-point verification statistics. The fuzzy verification technique is used by Grid-Stat to produce neighborhood verification statistics. This approach relaxes the normal requirement to verify at specific grid point locations and uses a neighborhood of points surrounding the forecast/observation pairs. Thus, a “close” forecast is given some credit (Ebert, 2008).[3] The spatial verification capability is called Method for Object-based Diagnostic Evaluation (MODE) which provides quantitative assessment of the degree to which forecast objects match the equivalent objects rendered from a gridded observation data set (Davis, 2009).[4] The statistical output from MODE generated using high resolution WRF can then be used to show its skill in forecasting objects. This could be compared to the same output from MODE generated using the NAM model forecasts. The difficulty in achieving this goal arises when attempting to acquire the gridded observation dataset for continuous field meteorological variables, which MODE requires. This gridded data set must have the same horizontal resolution as the WRF, which for future assessments will be 1 km or less.

Conclusions

The statistical results presented here are from the 31-day WRF and NAM model runs conducted over complex, mountainous terrain in Utah; thus, any conclusions drawn from the results are limited to this one environment. The case studies, conducted during the late spring and early summer of 2011, characterize the overall WRF and NAM model performance based on error statistics. These statistics were generated by comparing the forecast meteorological variable values interpolated from the model grid to the same value from a point-based observation. The WRF 1-km and 3-km models were compared with the 12-km NAM forecast output.

The comparison of error statistics for the 31-day study period suggests that while the majority of error statistic comparisons show that the WRF outperforms the NAM statistically, the values of these errors are not significant from an operational perspective, and the value added of the high resolution WRF over the NAM initialization forecasts is indeterminate.

There are significant errors in the forecasts for all three models of mean sea level pressure and row mean surface wind direction which merit further investigation as to the source of these errors.

Analysis of the upper-air forecast errors show no significant difference between the models. Any differences tend to occur at the lowest levels in and near the surface boundary layer.

[3] Ebert, E. Fuzzy Verification of High-Resolution Gridded Forecasts: a Review and Proposed Framework. *Meteorological Appl*. **2008**, *15*, 51−64.

[4] Davis et al. The Method for Object-Based Diagnostic Evaluation Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program. *Weather and Forecasting* **2009**, *24*, 1252−1267.

Recommendations

The MET has proven to be a powerful means assessing the accuracy of the WRF model. The automation effort of the three components of MET (Point-Stat, Grid-Stat, MODE) should be continued.

Further studies are needed which incorporate assessments of WRF performance using the traditional error statistics, combined with neighborhood and spatial verification techniques available from MET. The addition of information on spatial errors may enable more conclusive determinations of how models compare in their performance and the value added of one model over another.

1. Introduction

The atmosphere is a non-linear system with intertwined feedback loops. Consequently, accurate weather forecasting has proven to be a daunting task, even for state-of-the-art Numerical Weather Prediction (NWP) models. The models are typically segregated or "nested" according to space and time resolutions, such as synoptic, mesoscale, and microscale. The synoptic scale refers to weather systems that have a horizontal length scale of the order of 1000 km and typically pass over a given point in a period of one or two days (Huschke, 1959). According to Orlanski mesoscale includes weather systems of spatial scales from 2 km to 2000 km and temporal periods of 6 h to 24 h (Orlanski, 1975). Microscale meteorology is the study of short-lived atmospheric phenomena smaller than mesoscale, with about 1 km or less horizontal scale and a temporal period of seconds to a few hours (Wikipedia, 2012). The Army is mainly interested in weather phenomenon scales from the meso-gamma (2−20 km) to the microscale. Additionally, the Army focuses on the weather conditions in the boundary layer, a highly variable layer near the surface of the earth that changes frequently and diurnally, based on the time of day and synoptic atmospheric conditions. Forecasting in the boundary layer is challenging due to the interaction of the atmosphere with terrain, vegetation, buildings, and bodies of water.

The WRF model is a mesoscale numerical weather prediction system intended for operational forecasting and atmospheric research needs. The model was developed and improved by a collaborative partnership of the National Center for Atmospheric Research (NCAR), the National Oceanic and Atmospheric Administration (NOAA), the National Centers for Environmental Prediction (NCEP), the NOAA Global Systems Division, the Air Force Weather Agency (AFWA), the Naval Research Laboratory, the University of Oklahoma, and the Federal Aviation Administration. The Army, through the United States Air Force, has applied WRF to meet Army operational and research requirements. The U.S. Army Research Laboratory (ARL) has ongoing efforts to adapt, enhance, validate, and operate the WRF. A goal is to tailor the Advanced Research version of the Weather Research and Forecast (WRF-ARW) NWP model and couple it with an observation nudging, four-dimensional data assimilation (FDDA) approach. This model can be used for such projects as a high-resolution "nowcasting" tool (WRF-FDDA), a high-resolution input used in boundary layer/urban meteorological models and artillery trajectory simulators, a modeling tool to support the Army, a means for developing surface sensor placement strategies, and a test bed to investigate the potential value of sub-km modeling to the Army.

In short, this study attempts to answer the question: are the research efforts to improve the forecast accuracy at the finer scales worth the time, energy, and resource investment?

2. Methods, Assumptions, and Procedures

2.1 MET and the Automation of the Model Assessment Process

The Model Evaluation Tools (MET) is a set of verification tools developed by the WRF Developmental Testbed Center (DTC) for use by the numerical weather prediction community, especially users and developers of the WRF model, to help them assess and evaluate the performance of the models (National Center for Atmospheric Research, 2009).

The three main statistical analysis components of the current version of MET are Point-Stat, Grid-Stat, and the Method for Object-based Diagnostic Evaluation (MODE).

The Point-Stat tool is used for grid-to-point verification, or verification of a gridded forecast field against point-based observations (i.e., surface observing stations, rawinsondes, and other point observations). It provides forecast verification scores for both continuous (e.g., temperature) and categorical (e.g., rain) variables with associated confidence intervals. Confidence intervals take into account the uncertainty associated with verification statistics due to sampling variability and sample size limitations.

The Grid-Stat tool produces verification statistics when a gridded field is used as the observational dataset. Like the Point-Stat tool, it also produces confidence intervals. The Grid-Stat tool uses a fuzzy verification technique which employs a neighborhood verification method of relaxing the requirement to verify the forecast at a specific point and allows the verification to occur over a spatial window or neighborhood of surrounding points. The user can choose the size of the square search window within which the forecast event can fall and still be considered a “hit” (National Center for Atmospheric Research, 2009). When using Grid-Stat, severe limitations exist due to the lack of a suitable independent gridded observation dataset. There was no source for gridded observational data in our Utah study.

The MODE tool also uses gridded fields as observational datasets, defining objects in both the forecast and observation fields. This technique employs the use of spatial verification to attempt to assess model performance not captured by traditional, grid-to-point verification statistics. Spatial verification provides a quantitative assessment of the degree to which forecast objects match the equivalent objects rendered from a gridded observation data set. The statistical output from MODE can then be used to show the skill of the high resolution WRF in forecasting objects which could be compared to the same output from MODE generated using the lower resolution NAM model forecasts. This quantifies the closeness of the forecast object to the observed object through the use of several attributes which characterize objects. The difficulty in using MODE arises when attempting to acquire the gridded observation dataset for continuous field meteorological variables which it requires. This gridded data set must have the same horizontal resolution as the WRF, which for future assessments will be 1 km or less.

The ARL Model Assessment Project will eventually utilize all three components of the MET. The initial phase started in 2009 and continued through 2011, focusing on the use of the MET Point-Stat tool with a domain centered on the Dugway Proving Ground (DPG), Utah (figures 1 and 2).

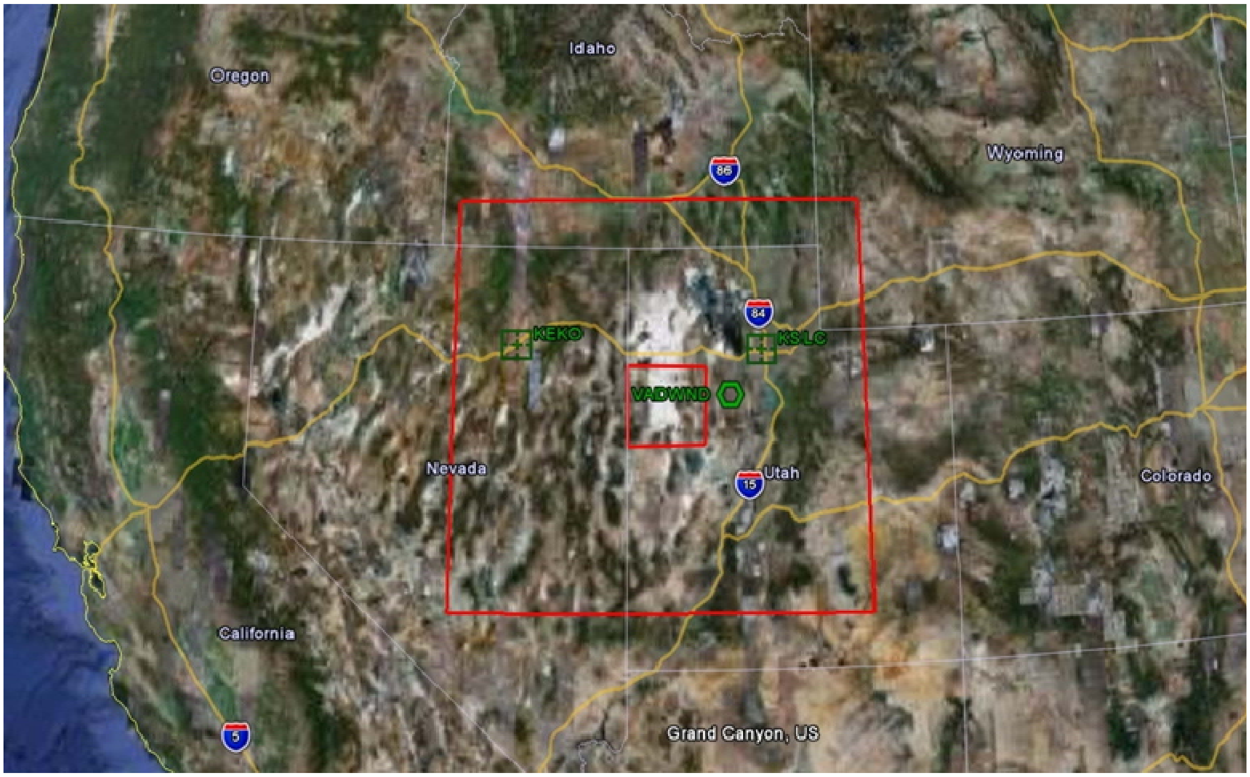

Figure 1. DPG Domains 1 and 2.

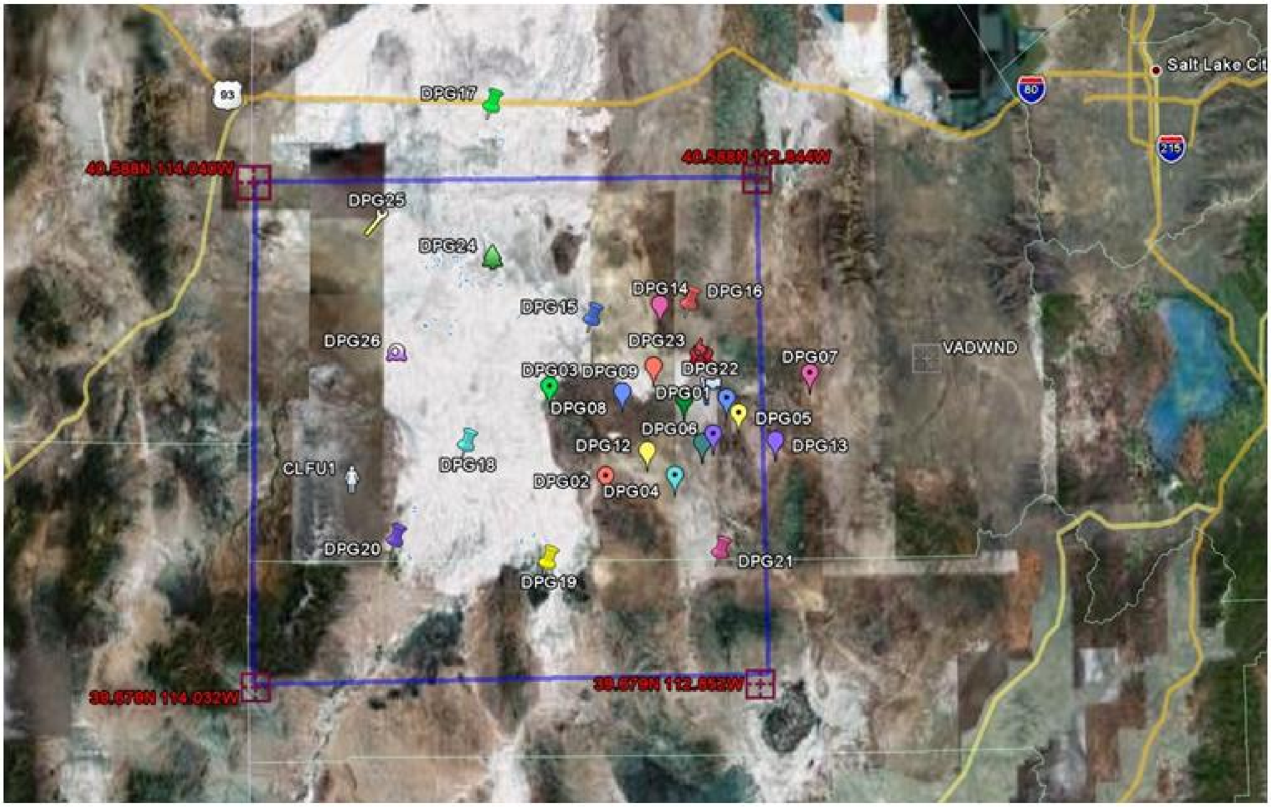

Figure 2. Expanded view of Domain 2.

A cornerstone study conducted in 2009 incorporated the use of the MET for the first time at ARL and enabled more comprehensive statistical evaluations of the WRF model, using a forecast grid-to-point observation model verification method (Sauter, 2009).

In a follow-on assessment, those previously-developed methods were used as the foundation for a more robust implementation of MET accomplished in 2010, which included a twentyfold increase in Meteorological Assimilation Data Ingest System (MADIS) observational data. The MET process was also expanded to generate Mean Error (ME), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) statistics based on 140 WRF runs, which used seven different parameterization settings and two different WRF resolutions, for 20 case study days during the period March 2009 to July 2010. The sequential tasks of post-processing the WRF output data, acquiring the required observation data, performing data format conversions, and running Point-Stat and Stat-Analysis routines on several different computer platforms required the coordination of over 100 Unix scripts.

This enhanced set of verification statistics, generated using grid-to-point evaluations, revealed that the errors varied depending on the case study day which may have been related to the particular weather situation occurring on those days. This case-study approach may have skewed the results towards model performance in these specific types of weather conditions. Another conclusion drawn from this extensive set of statistics was that the different parameterization settings had little impact on the value of the error statistics, with the possible exception of one of the boundary layer parameter settings. All settings essentially produced the same error values. In addition, the errors for the 1-km and 3-km horizontal resolution WRF versions were statistically the same (Raby, 2011).

2.2 Model Performance During a Continuous Evaluation Period

To address the possible Biases introduced by the earlier weather-driven case study approach, the 2011 study emphasized running and evaluating the WRF model during a continuous study period. During the 31-day period from 27 May to 26 June 2011, the WRF was run each day. To establish a benchmark comparison for WRF performance, output of the North American Mesoscale (NAM) model, a variant of the WRF model, was acquired and assessed. The NAM model was run over the same domains as the WRF model, and served as the initialization grid for the WRF, but at a lower horizontal resolution of 12 km. Assessing the errors of the NAM model would then provide a basis for WRF error comparison to show if there was value added in running the higher resolution WRF.

Unlike the previous study of 20 distinct case study days, the approach used for this assessment was to aggregate the statistics generated, using the grid-to-point method over 31 consecutive model runs. For surface meteorological variables, these comparisons were accomplished using two different aggregations. The first was to aggregate the errors over all the runs and output single statistics, which characterize the model errors for all 24 h of the model run over the entire 31-day study period. For these statistics, the 95% bootstrap confidence intervals were extracted to capitalize on the extensive sample size in order to obtain an assessment of how the three models compared and if the observed differences are statistically significant. The second was to aggregate the errors over all the runs and output the statistics by forecast lead time to see how the models compared over three consecutive forecast periods, starting at 1200 Coordinated Universal Time (UTC) and ending at 1800 UTC. For upper-air meteorological variables, the model comparisons were accomplished by aggregating the errors generated from the differences between data collected from two rawinsonde upper-air observation sites and the model forecasts at various level in the atmosphere.

2.3 Case Studies

The WRF model was run with an outer nest over Utah and portions of surrounding states and a single inner nest centered over DPG. These domains are shown by the red rectangles on the first map (figure 1). The outer nest (Domain 1) is a square having 546 km on a side and was run with a grid spacing of 3 km, while the inner nest (Domain 2) is a square having 102 km on a side and was run with a grid spacing of 3 km and 1 km. For the 3-km WRF run over Domain 2, the results were interpolated onto the 1-km inner nest grid space. In the case of the 1-km WRF run over Domain 2, the results populated the 1-km inner nest grid space.

WRF version 3.2.1 was used in this study. The parameters used for this modeling study were as follows:

- WRF single-moment microphysics scheme
- No cumulus parameterization scheme
- 3:1 grid space (km) to advected time step ratio
- Dudhia short-wave radiation
- Rapid Radiative Transfer Model (RRTM) long-wave radiation
- Noah Land Surface Model
- Yonsei State University (YSU) Planetary Boundary Layer and surface layer schemes
- Terrain slope/shadow option used
- Internal time step = 9 seconds
- 60 vertical levels

WRF runs were initialized at 0600 UTC with output generated every hour from 0 to 24 h for surface and upper-air meteorological variables each day during the period 27 May through 26 June 2011.

WRF output, interpolated from model sigma terrain-following coordinates onto pressure-level surfaces was generated with the WRF Post Processor Version 3 (WPPV3), and those values were compared to point observations including surface, upper air, and aircraft data.

The NAM model utilized had a 12-km horizontal resolution grid. Archived output for each day was downloaded from the NCAR Computational and Information Systems Laboratory Research Data Archive. The output files were extracted, then converted from Gridded Binary format (GRIB2) to GRIB1 format and the 24-h forecast from 0600 UTC base time was retained. The NAM output contains forecasts of surface and upper-air meteorological variables at 3-h intervals. The output was compared to point observations including surface, upper-air, and aircraft data over Domains 1 and 2, by applying a Point-Stat masking region which restricts the scoring to those areas.

All the observations were within 21 min before or after the model valid time on the hour. The Meteorological Terminal Aviation Weather Report (METAR) observations were obtained from the NCEP PrepBUFR files for Domain 1. Approximately 20 to 25 PrepBUFR surface station observations were available each hour for Domain 1, with occasionally only one surface observation within Domain 2. The PrepBUFR observations also include two upper-air soundings located at Salt Lake City, UT (KSLC) and Elko, NV (KEKO) and sporadic aircraft observations. MADIS mesonet data were added to the PrepBUFR METAR observations increasing the number of observations to approximately 500. These also included approximately 25 surface station observations within Domain 2, as shown in figure 2. These Domain 2 mesonet surface data are primarily over DPG, and their quality control was considered acceptable after being subjected to MADIS Meteorological Surface Quality Control Level 3 checking (MADIS Meteorological Surface Quality Control, 2012). No upper-air soundings were available within Domain 2 for this study.

This report documents the ME or Bias, MAE and RMSE error statistics and uses them to characterize model performance. Nonparametric confidence intervals at the 95% level for the above statistics were computed using the bootstrap method provided by MET. This method uses samples of the verification statistics to infer the uncertainty information for the entire set of forecast-observation pairs collected. This avoids the assumption that the distribution of the error statistics is normal and assumes that the inference made from the samples is representative of the true population distribution (National Center for Atmospheric Research, 2009). As a matter of convenience, results were noted to two decimal places even though the data were not significant to that degree of accuracy. Errors were omitted for any observed wind speed less than 1 m/s. MET calculates wind direction errors in two different ways:

1. For the “ROW_MEAN_WDIR” line, for each forecast valid time, the mean forecast wind direction, mean observation wind direction, and the associated error are computed for each forecast-observation wind component (U and V) vector difference. Then the means are computed across each of these forecast wind directions, observation wind directions, and their errors. MET Point-Stat computes only ME and MAE statistics for “ROW_MEAN_WDIR”.

2. For the “AGGR_WDIR” line, all the wind component forecast vectors are summed. Then the wind component observation vectors are summed. The vector difference between these two summed (aggregated) vectors provides an aggregated difference from which, the mean forecast wind direction, observation wind direction, and the associated error are computed and written out. MET Point-Stat computes only the ME statistic for “AGGR_WDIR”.

Both wind direction errors are included in this report. Note: Bias values near 180° are misleading since they are actually very close to a 0° Bias.

3. Results and Discussion

3.1 Extraction and Depiction of Case Study Results

The depiction of statistical results for the 31-day study period was achieved by extracting the statistics from the Stat-Analysis files. For this report, the results for surface and upper-air meteorological variables over all study days were calculated. The surface variable results were differentiated by forecast hour and by model horizontal resolution. The upper-air results were differentiated by forecast hours 0000 and 1200 UTC and by model resolution.

Tabular results were produced by importing extracted database files directly into MS Excel worksheets. The line and scatter charts were created using the MS Excel chart tools. The line chart option compared surface variable statistics for the three model resolutions and showed how the statistics varied by forecast hour for the three model resolutions. The scatter chart option was used for plotting the upper-air statistics for two model resolutions. If the number of forecast-observation pairs for any variable was less than two, the error statistic was not plotted but did appear in the tabular data.

The vertical scale for some of the plots had to be adjusted, to exploit the full range of the error statistic values; thereby revealing more detail about the error behavior as a function of forecast hour.

3.2 Characterization of Case Study Results

The results of the study were characterized in two ways. One way was a quantitative determination of the difference significance between the error statistics of each of the three models. The purpose of this calculation was to decide whether the error statistics were different enough to make the statement that one model had a larger error than another model, or whether that difference was insignificant due to the amount of uncertainty in the error values themselves, owing to the limited number of samples used to produce the error values. This assessment is referred to as "statistical" significance, for this study. Statistical significance was determined by comparing plotted error values for each model and their associated uncertainty in the form of error bars. The error bars were calculated by using the 95% bootstrap confidence interval statistics produced by the MET. If the range of error encompassed by the error statistics and its associated error bars for a given model did not overlap that of another model, then it was concluded that the error statistics for the two models were significantly different from each other. This means that the value of the error statistic for the model, whose error was lowest, was actually lower by comparison and not just randomly lower, because of the error generated by the limited amount of sampling. However, if the range of error for a given model DOES overlap that of another model, it cannot be said that there is no significant difference in the error statistics of those two models (Cornell University Statistical Consulting Unit, StatNews#73, 2012).

The other way that the results of the study were characterized was by a subjective assessment of the "operational" significance of the values of the error statistics themselves. This characterization was the result of a judgment made by competent meteorologists familiar with how forecasts of meteorological variables impact military operations. An example of how this characterization was applied is as follows:

It may be observed that the RMSE values for 2-m level air temperature for three models A, B, and C are 2.0, 2.5, and 2.9 °K respectively. While one might characterize the difference between Model A and Model C as being statistically "significant", because their error bars did not overlap, operationally, the impact of a temperature forecast error of 2.0 °K is effectively no different than a temperature forecast error of 2.9 °K. Thus, the judgment in this case is that the performance of these two models is the same since the difference in the impact of their errors on operations is nil.

3.3 Comparison of Overall Surface Meteorological Variable Errors for the Three Model Resolutions

Figures 3 through 5 are plots that show the surface temperature (2-m) errors for Domain 2. Figure 3 shows the comparison of the RMSE errors between the 1-km WRF, the 3-km WRF, and the 12-km NAM. Figure 4 displays the MAE between the three models, while figure 5 shows the Bias error over the 31-day study period.

The RMSE spread, including confidence intervals was less than 0.3 °K between the three models, with the 1-km WRF (blue) most favorable (lowest RMSE) and the 12-km NAM (green) least favorable (highest RMSE). The MAE differences seen in figure 4 covered less than 0.3 °K between models with both WRF (1-km and 3-km) models having the lower MAE values. As seen in figure 5, the 1-km WRF and 3-km WRF (red) both overforecasted 2-m temperature, while the 12-km NAM slightly underforecasted the temperature over the 31-day study. Although the 1-km WRF RMSE and MAE were statistically significantly less than the NAM, the absolute value of the NAM Bias was less than both WRF models as the NAM error was about −0.1 °K and the WRF errors were between 0.8 °K and 0.9 °K. It is interesting to note that the Bias results for the 3-km WRF for Domain 1 (not shown) indicate a reversal from overforecasting to underforecasting. These differences in error statistics are not considered significant from an operational standpoint.

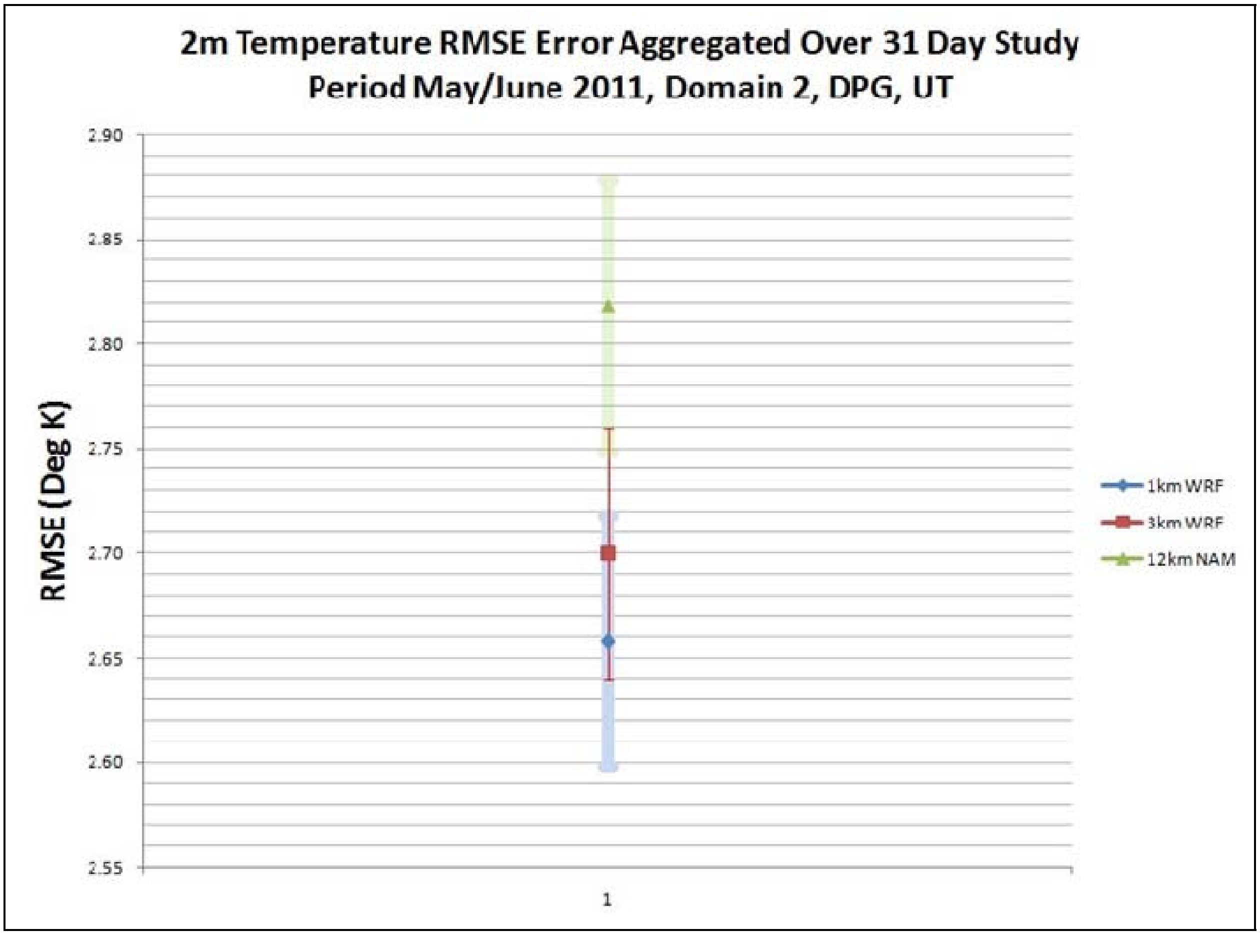

Figure 3. Comparison of the 2-m air temperature RMSE statistic for Domain 2.

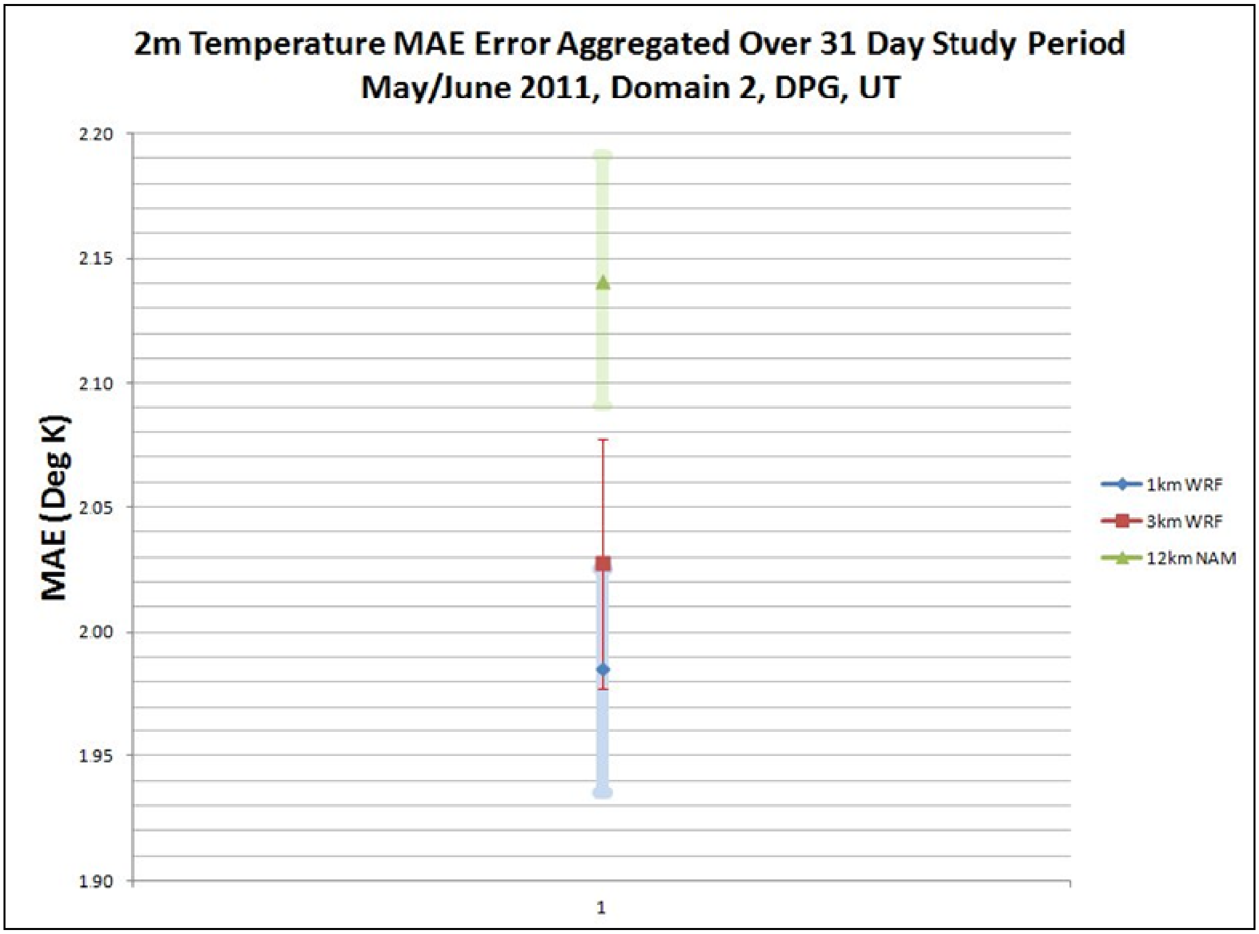

Figure 4. Comparison of the 2-m air temperature MAE statistic for Domain 2.

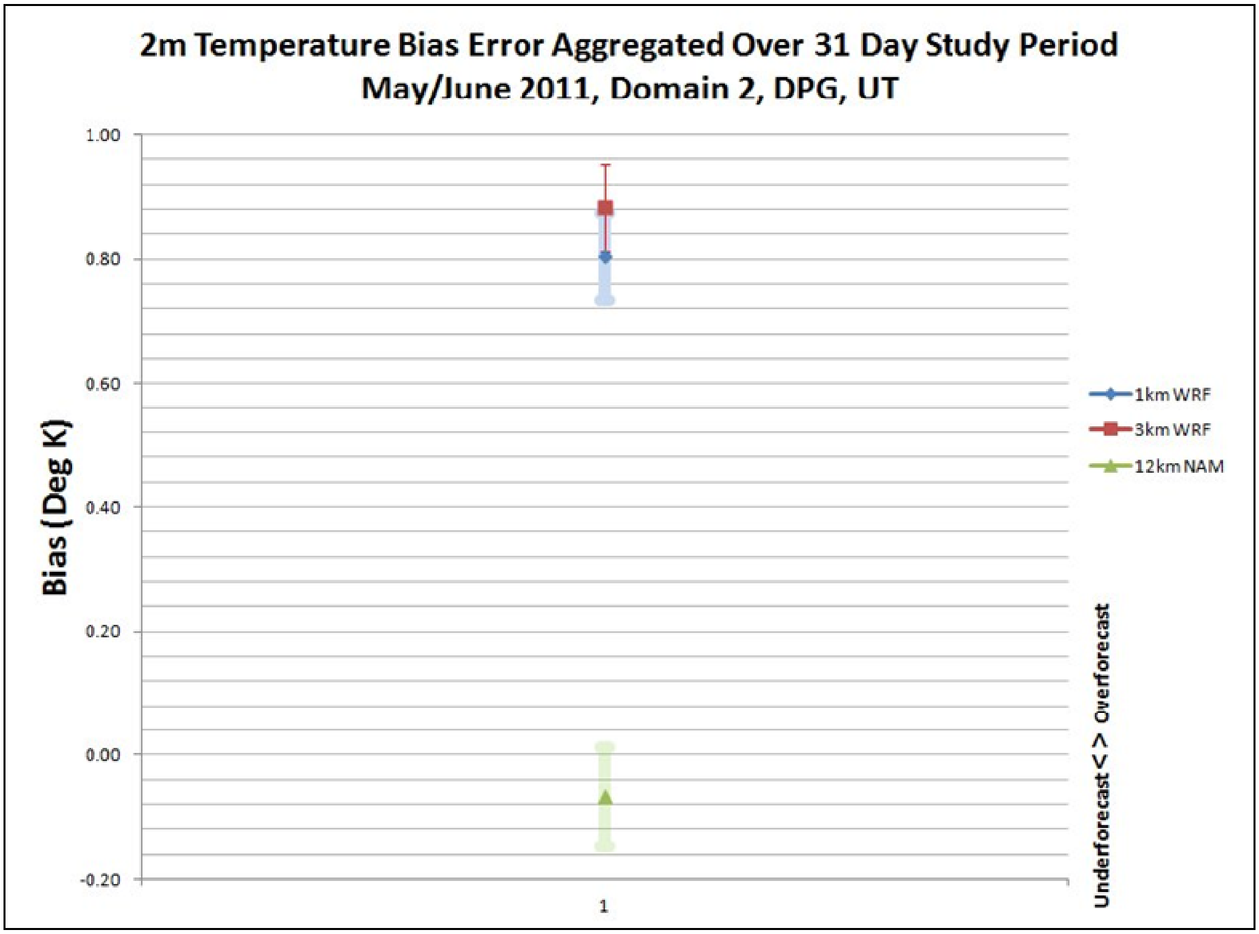

Figure 5. Comparison of the 2-m air temperature Bias statistic for Domain 2.

Figures 6 and 7 are plots that show the surface (2-m) dew point temperature errors for Domain 2.

Figure 6 displays the RMSE for the surface dew point between the models while figure 7 shows the Bias. The difference in the RMSE between the WRF and the NAM was about 0.6 °K and the WRF outperformed the NAM with statistical significance. The 1-kmWRF and 3-kmWRF errors grouped closely together, and were statistically significantly lower than the 12-km NAM by 0.5 °K (including error bars). The 1-kmWRF and 3-km WRF both exhibit a dry Bias (underforecast), and the 12-km NAM shows a moist Bias (overforecast). In contrast with temperature, statistically the WRF outperforms the NAM but, judging from an operational perspective, these differences are not considered significant.

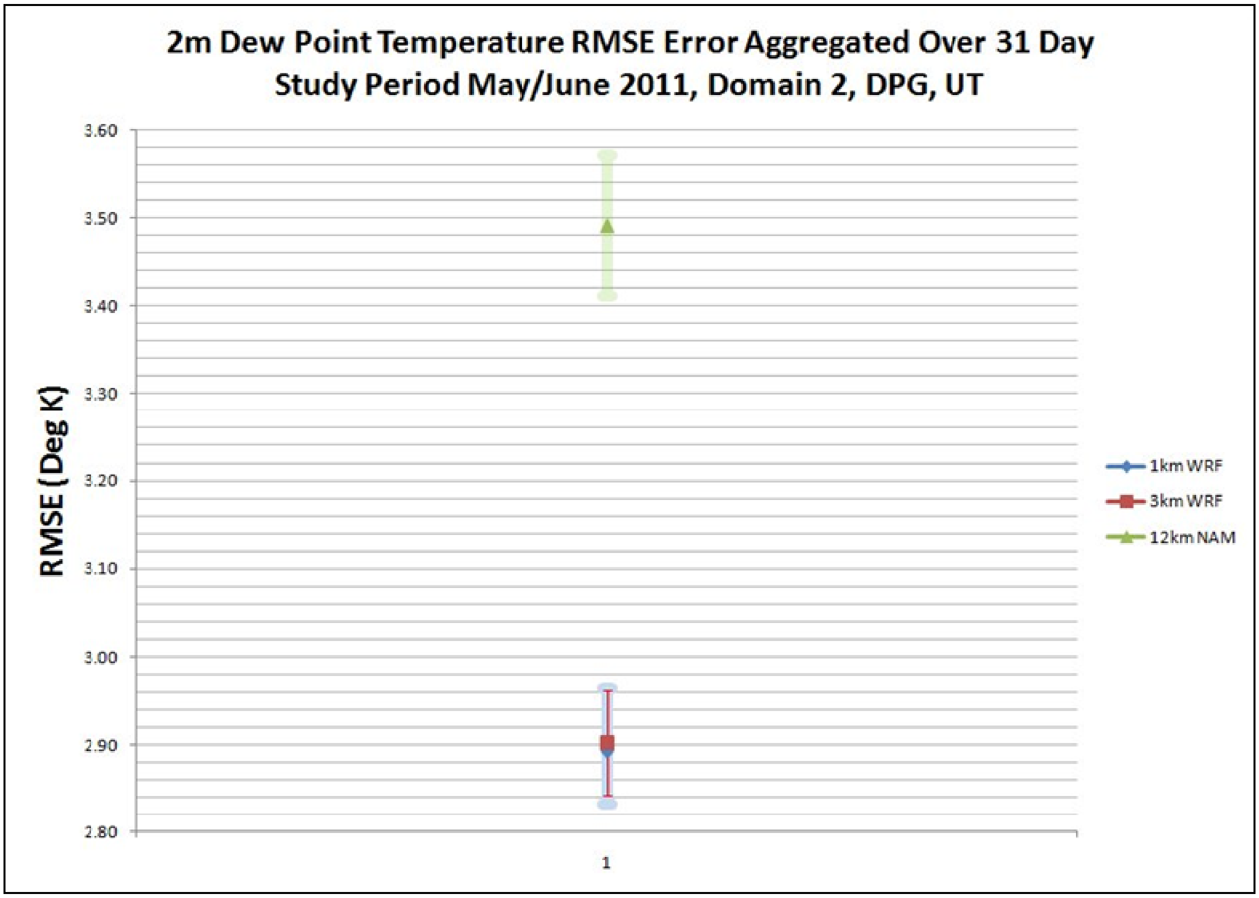

Figure 6. Comparison of the 2-m dew point temperature RMSE statistic for Domain 2.

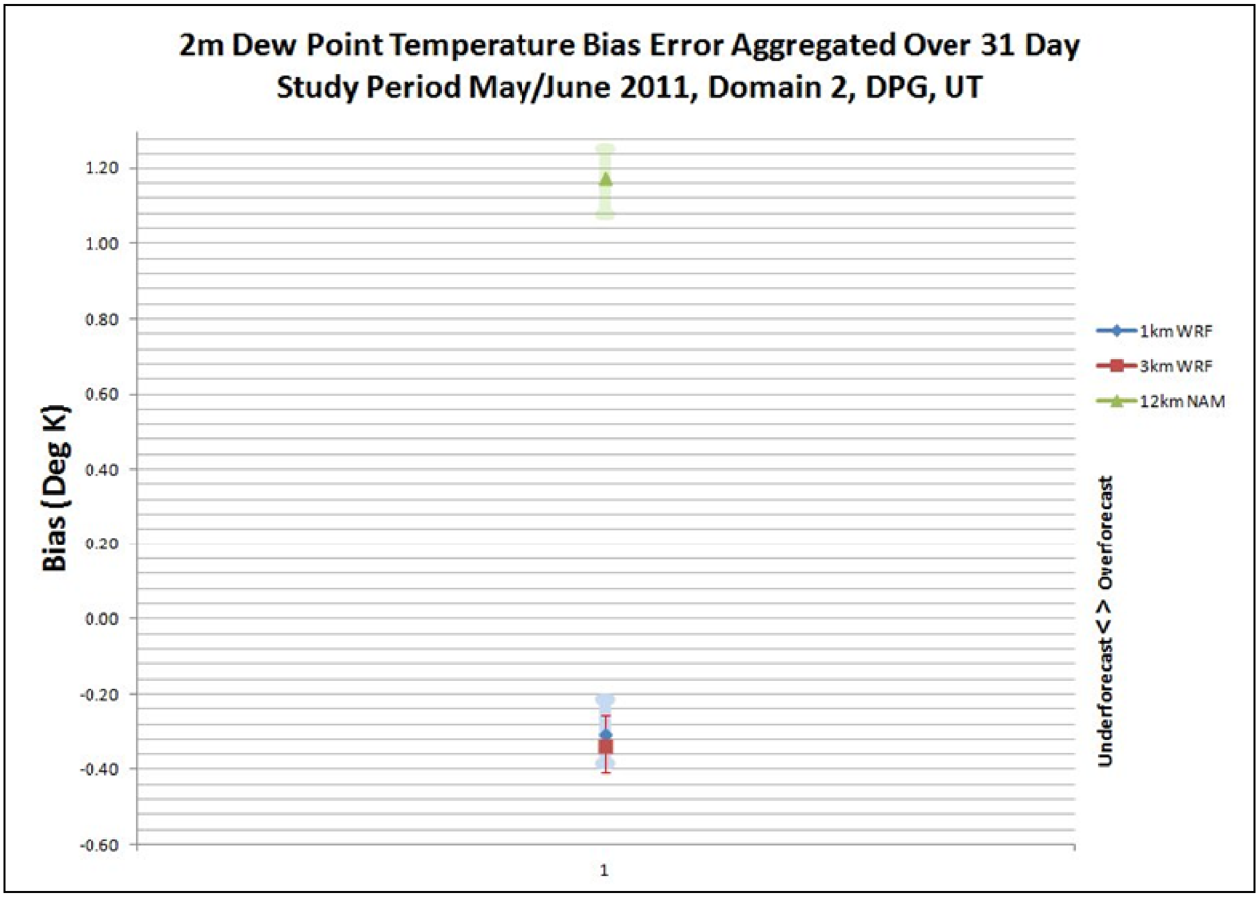

Figure 7. Comparison of the 2-m dew point temperature Bias statistic for Domain 2.

Figures 8 through 10 are plots that show the mean sea level pressure errors for Domain 2, where figure 8 is the RMSE, figure 9 the MAE, and figure 10 the Bias. The 1-km WRF, 3-kmWRF, and 12-km NAM RMSE errors overlapped one another with a spread of less than 0.3 hPa. The MAE results were consistent with the RMSE, with a slightly smaller range of values. All models exhibited a notable tendency to underforecast the mean sea level pressure. No statistically significant differences were noted between the models, but the magnitude of their errors is considered significant. As a comparison, error statistics for surface mean sea level pressure from the 15-km WRF (provided by AFWA) averaged over Continental United States (CONUS) (figure 11) and Hill AFB, UT (figure 12) are presented (Air Force Weather Agency, 2011). The Hill AFB plot shows continuous traces of the forecast values of mean sea level pressure for each 6-h model run cycle in red, green blue and yellow and the observed pressure in black. In contrast to the errors presented in this study at Dugway, the 15-km results show RMSE and Bias errors on the order of 2.0 hPa or less with a slight negative Bias and slight positive Bias, such as the case at Hill AFB, UT (not shown). While there are some cases at Hill AFB that do display large errors in sea level pressure, most of the days in the June 2011 study at that location show smaller errors than seen in the current study. As with all comparisons, it should be noted that the ARL study is focused on 1-km and 3-km WRF output and not 15-km output as with the AFWA cases shown in figures 11 and 12. The source of this error over Dugway is not known and this needs to be investigated.

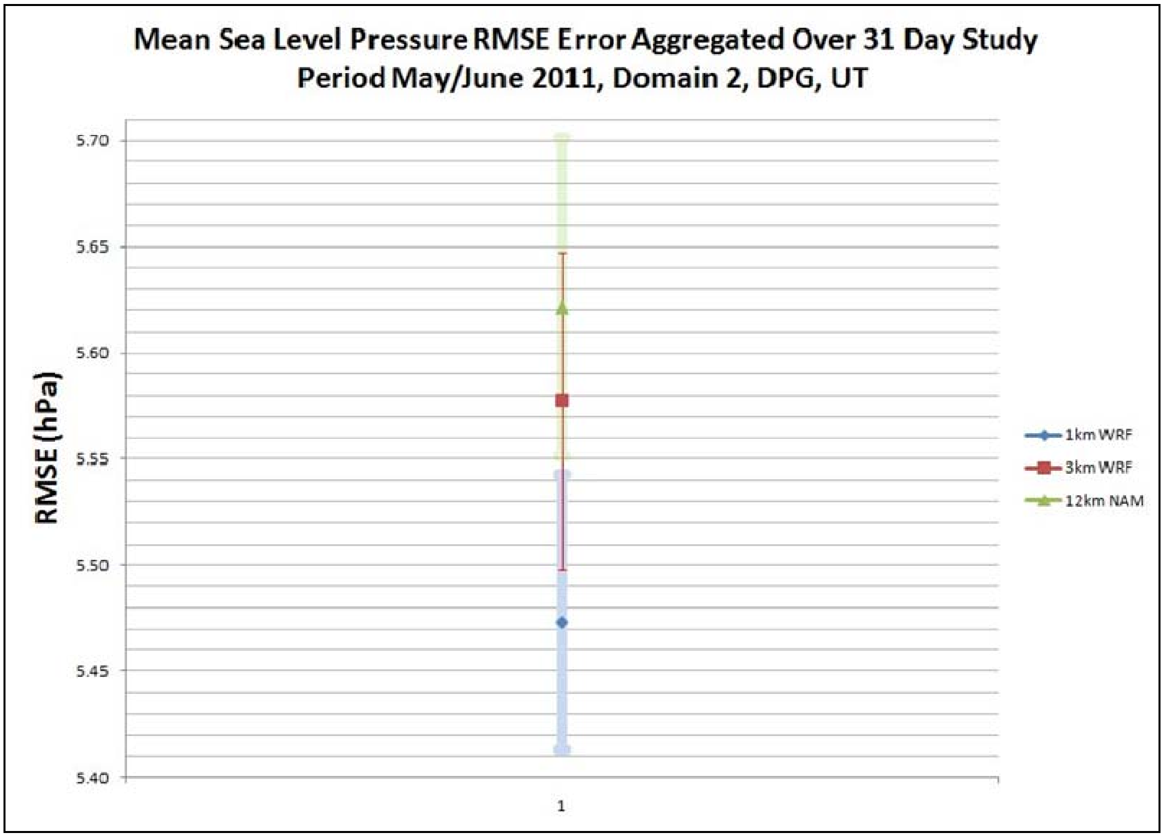

Figure 8. Comparison of the mean sea level pressure RMSE statistic for Domain 2.

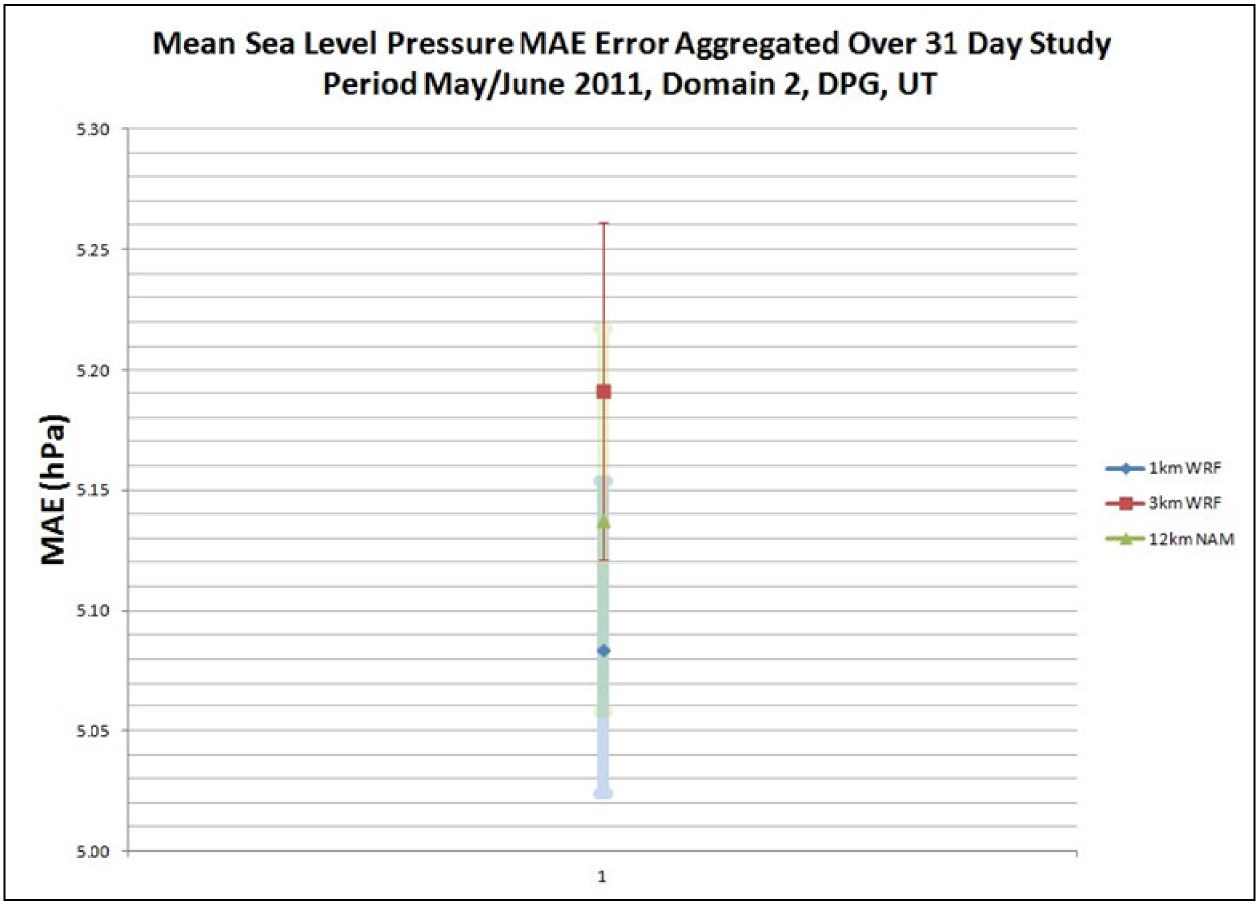

Figure 9. Comparison of the mean sea level pressure MAE statistic for Domain 2.

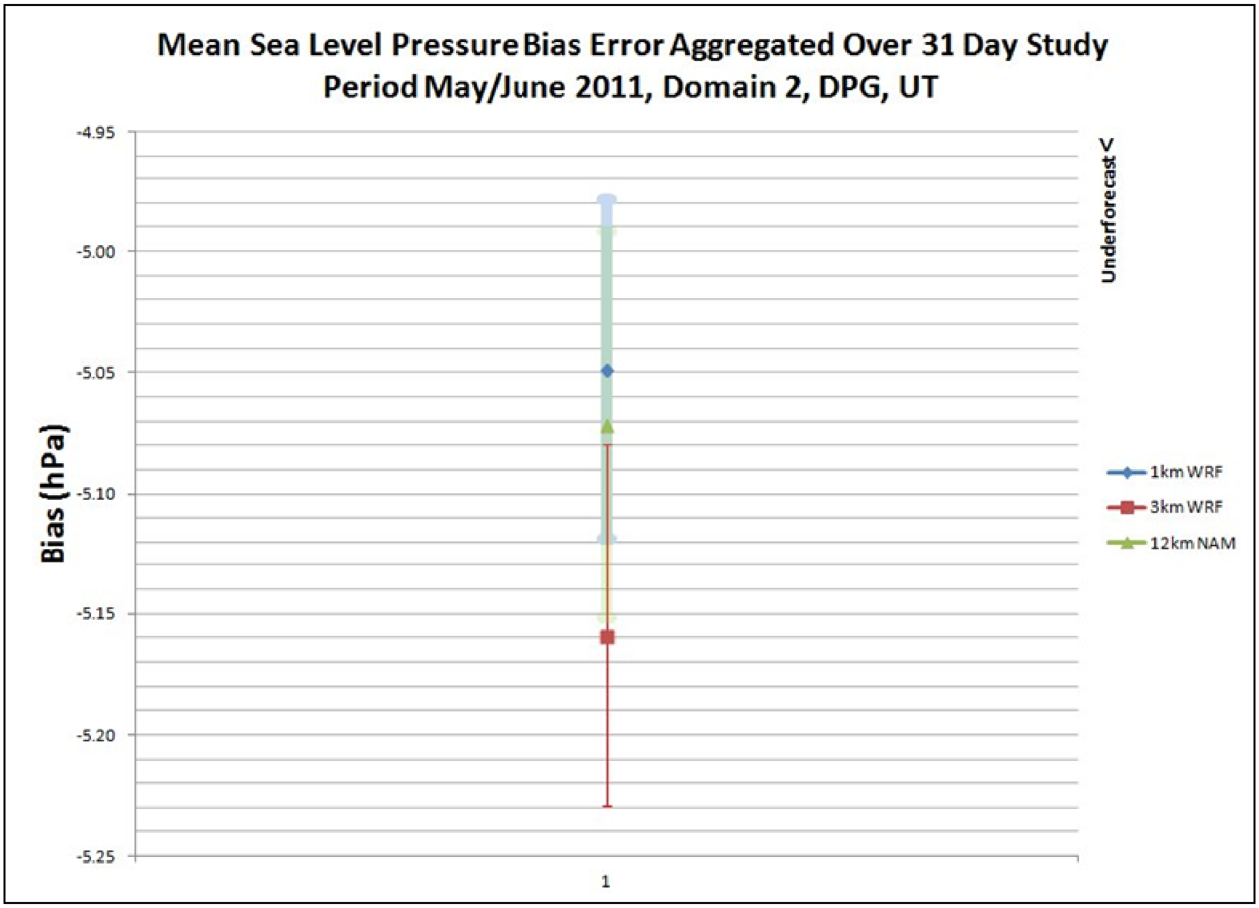

Figure 10. Comparison of the mean sea level pressure Bias statistic for Domain 2.

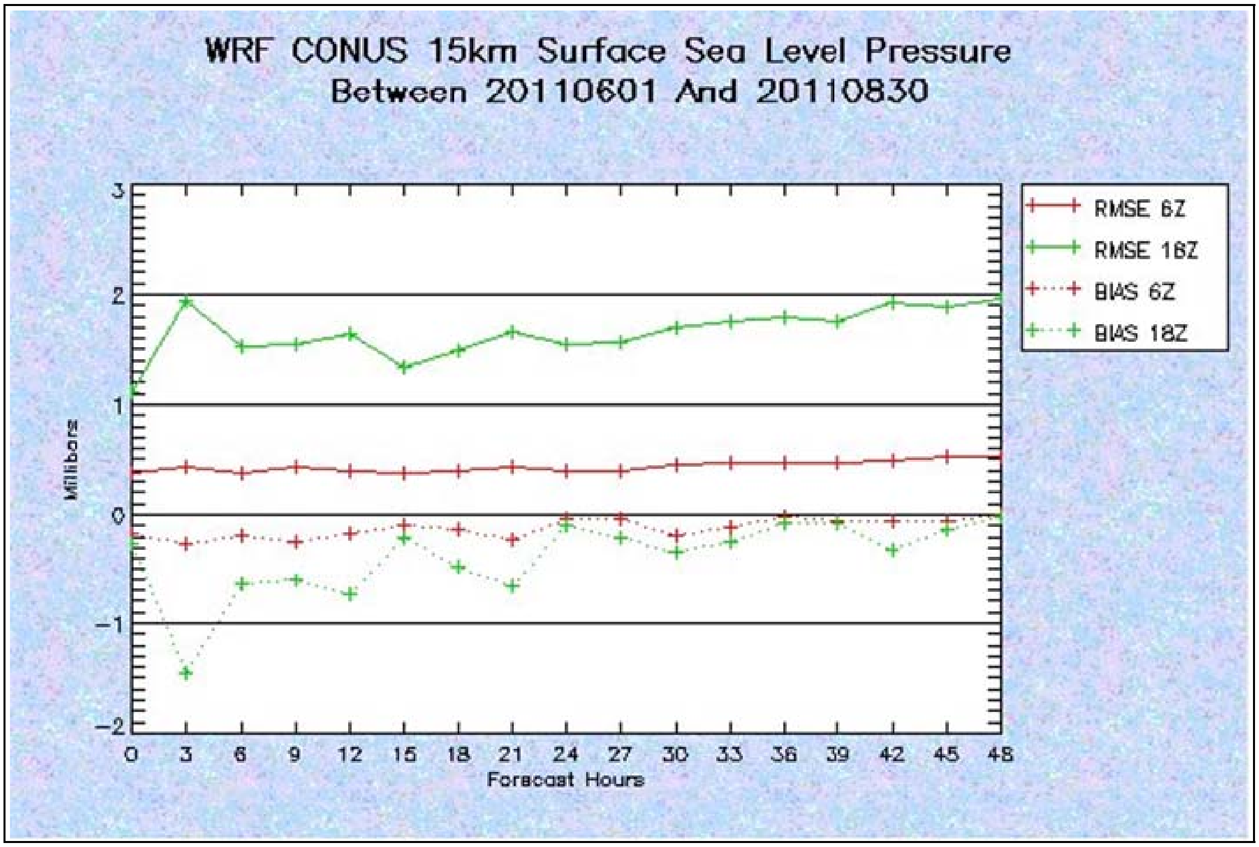

Figure 11. AFWA sea level pressure for CONUS RMSE and Bias statistics for 15-km WRF by forecast hour.

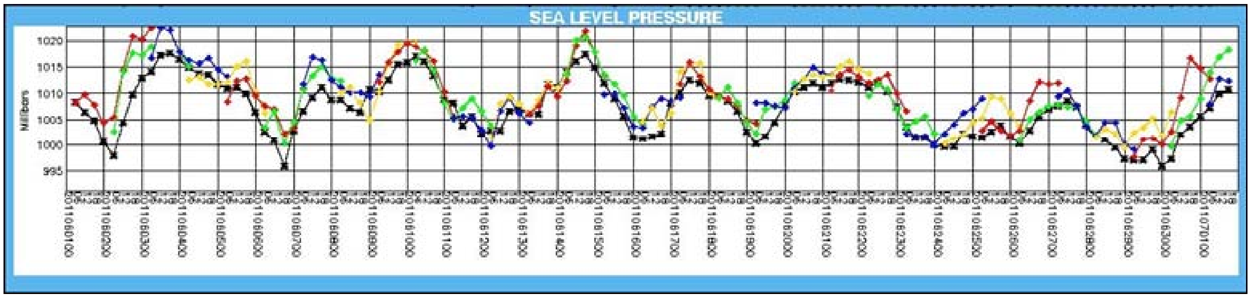

Figure 12. AFWA sea level pressure forecasts vs. observations for 15-km WRF for Hill AFB, UT for June 2011. The black trace is the observed pressure and the forecast pressure from the various forecast cycles are the other traces.

Figures 13 and 14 are plots that show the surface wind speed errors for Domain 2 where figure 13 shows the 10-m wind speed RMSE and figure 14 displays the wind speed Bias. All three models produced overlapping RMSE and MAE (not shown) results. The range of magnitudes of the error was less than 0.3 m/s. The Bias for WRF being less in absolute value than that of the NAM is statistically significant; however, all models underforecasted the wind speed. It is interesting to note that the Bias results for Domain 1 (not shown) comparing the 3-km WRF and the 12-km NAM show a reversal from underforecasting to overforecasting. None of the differences between models appear to be of operational significance.

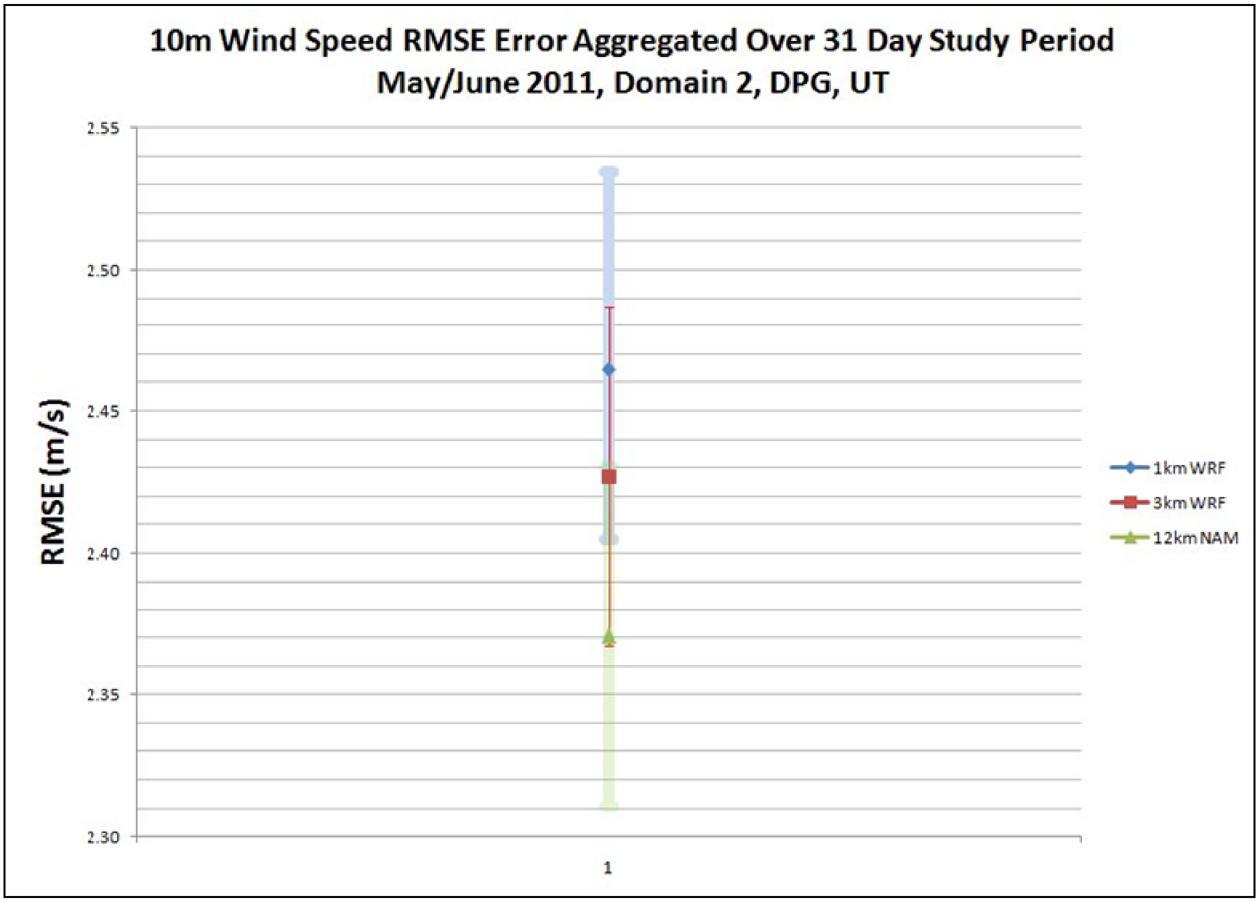

Figure 13. Comparison of the 10-m wind speed RMSE statistic for Domain 2.

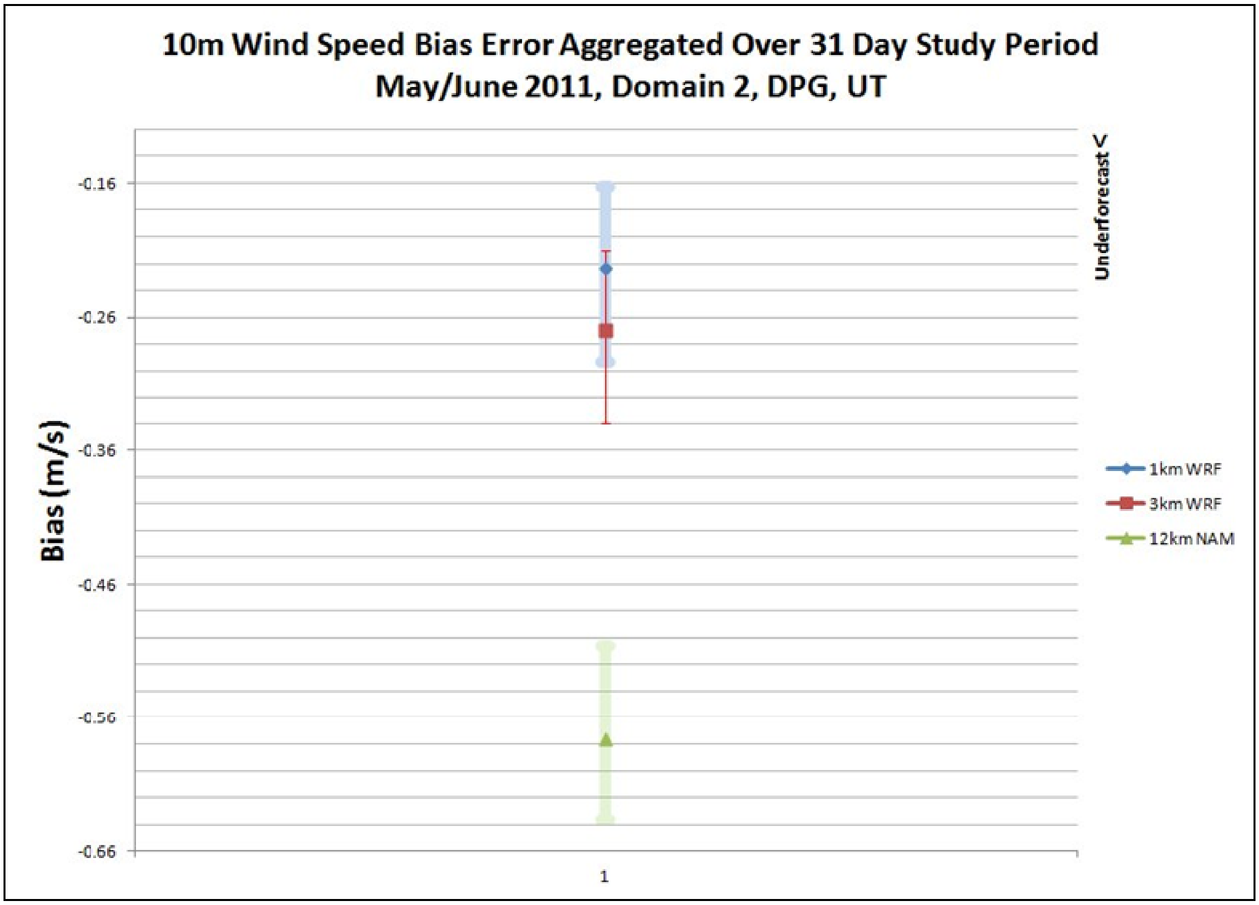

Figure 14. Comparison of the 10-m wind speed Bias statistic for Domain 2.

Figure 15 (MAE) and figure 16 (Bias) are plots that show the surface row mean wind direction errors for Domain 2. MET does not compute the RMSE statistic or the confidence intervals for row mean wind direction. The MAE results showed the 12-km NAM has less error than both the 1-km and 3-km-WRF; however, there is less than 1.4° separating the calculated magnitudes. Figure 16 shows that the Bias for the two WRF models is nearly identical, as indicated by their overlapping graphic icons. Both models underforecasted the wind direction where underforecasting would indicate backing of the winds and overforecasting would equate veering of the 10-m wind direction. The 12-km NAM overforecasted wind direction by about 14°. No significant wind direction differences were noted between models when investigating the MAE, but the magnitude of their errors is noteworthy. This appears to be one of the most interesting results of this study—the wind direction magnitude of the WRF 1-km and 3-km models are far less than the error of the 12-km NAM.

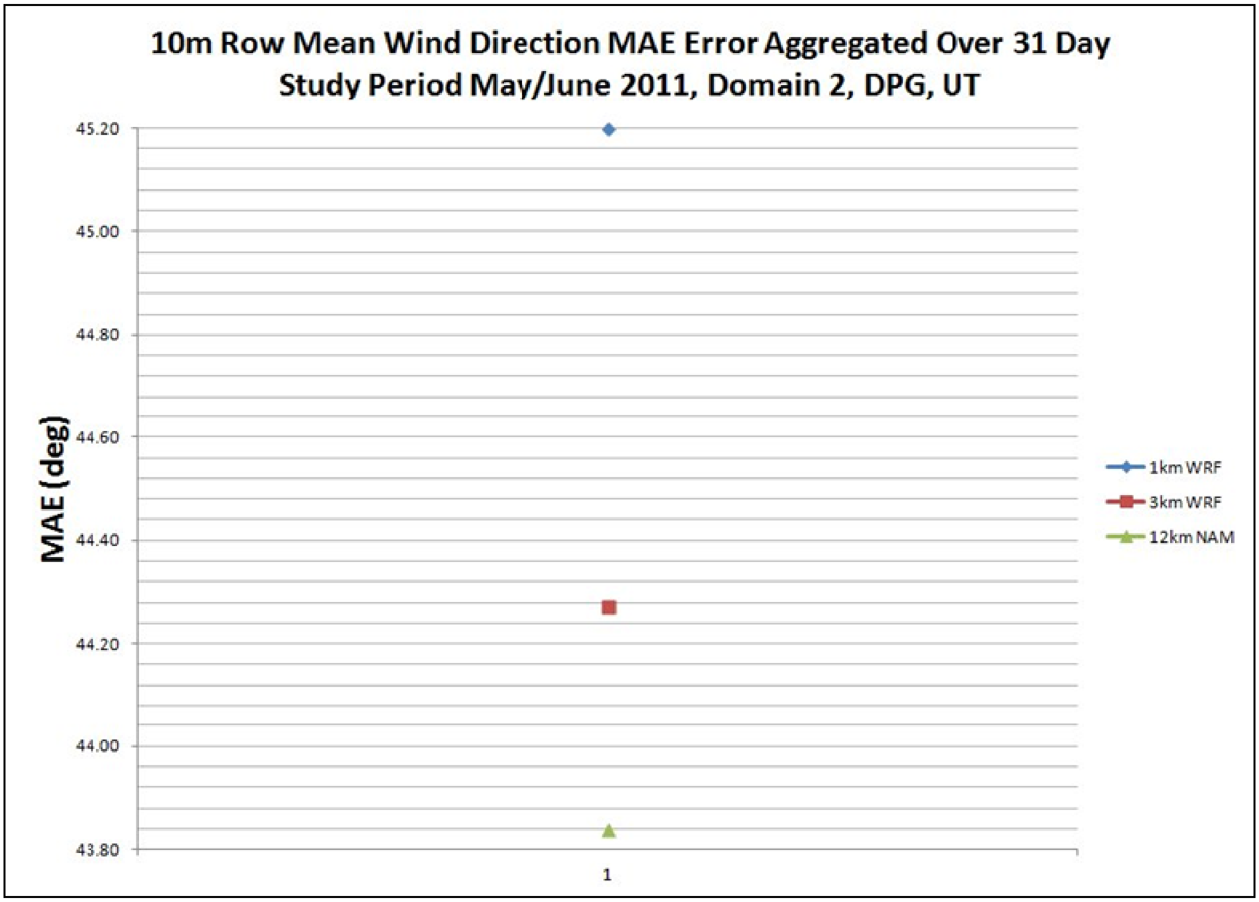

Figure 15. Comparison of the 10-m row mean wind direction MAE statistic for Domain 2.

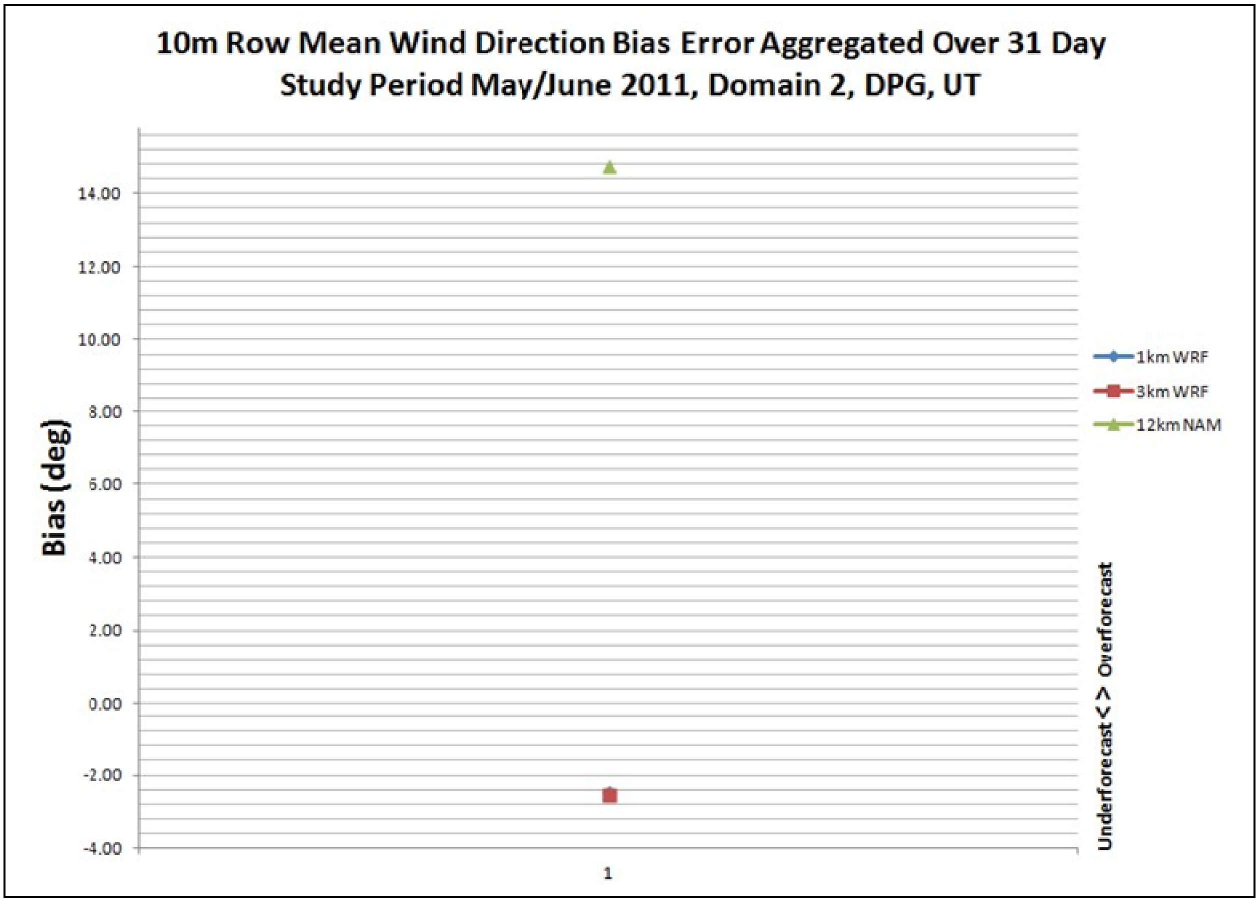

Figure 16. Comparison of the 10-m row mean wind direction Bias statistic for Domain 2.

The results for surface variables show that the errors are not significant from an operational perspective with the possible exception of mean sea level pressure and row mean wind direction. The errors for the two WRF models tend to group closely together and typically are not significantly different from each other. In the majority of cases, the WRF (1-km and 3-km) errors were statistically significantly lower than those of the NAM, but the magnitude of the error values from both models are not sufficiently different to demonstrate the value added of one model over the other.

Table 1 presents the surface results for all days and hours from the 31-day study in tabular form. The variables displayed are the temperature, dew point, RH, mean sea level pressure, along with wind direction and wind speed. Statistics examined are the RMSE, MAE, and Bias. Results shown are for both Domain 1 and Domain 2. These results show the outcome when the WRF and NAM models are compared using three categories to characterize all possible outcomes as follows:

1. Either the 1-km or 3-km resolution WRF error statistic (Bias, MAE and RMSE) had a statistically significantly smaller value (no error bar overlap in graphical comparison) than the value of the same NAM error statistic. This case is denoted as “**WRF < NAM**”.
2. Either the 1-km or 3-km resolution WRF error statistic (Bias, MAE and RMSE) was statistically equal in value (any error bar overlap in graphical comparison) to the value of the same NAM error statistic. This case is denoted as “**WRF = NAM**”.
3. Either the 1-km or 3-km resolution WRF error statistic (Bias, MAE and RMSE) had a statistically significantly larger value (no error bar overlap in graphical comparison) than the value of the same NAM error statistic. This case is denoted as “**NAM < WRF**”.

Table 1. Categorical analysis of overall surface meteorological variable errors for the three models.

Variable/Statistic/Domain	**WRF < NAM**	**WRF = NAM**	**NAM < WRF**
2m TMP /RMSE/2	X		
2m TMP /RMSE/1	X		
2m TMP /MAE/2	X		
2m TMP /MAE/1	X		
2m TMP /Bias/2			X
2m TMP /Bias/1	X		
2m DP /RMSE/2	X		
2m DP /RMSE/1	X		
2m DP /MAE/2	X		
2m DP /MAE/1	X		
2m DP /Bias/2	X		
2m DP /Bias/1	X		
2m RH /RMSE/2	X		
2m RH /RMSE/1	X		
2m RH /MAE/2	X		
2m RH /MAE/1	X		
2m RH /Bias/2			X
2m RH /Bias/1	X		
MSLP /RMSE/2	X		
MSLP /RMSE/1		X	
MSLP /MAE/2		X	
MSLP /MAE/1			X
MSLP /Bias/2		X	
MSLP /Bias/1			X
10m U-WIND/RMSE/2			X
10m U-WIND/RMSE/1	X		
10m U-WIND/MAE/2			X
10m U-WIND/MAE/1	X		
10m U-WIND/Bias/2		X	
10m U-WIND/Bias/1			X
10m V-WIND/RMSE/2		X	
10m V-WIND/RMSE/1		X	
10m V-WIND/MAE/2		X	
10m V-WIND/MAE/1	X		
10m V-WIND/Bias/2	X		
10m V-WIND/Bias/1	X		
10m WIND/RMSE/2		X	
10m WIND/RMSE/1		X	
10m WIND/MAE/2		X	
10m WIND/MAE/1	X		
10m WIND/Bias/2	X		
10m WIND/Bias/1	X		
10m RM-WDIR/MAE/2			X
10m RM-WDIR/MAE/1	X		
10m RM-WDIR/Bias/2	X		
10m RM-WDIR/Bias/1	X		
10m AG-WDIR/Bias/2	X		
10m AG-WDIR/Bias/1	X		
Category Totals	30	10	8

From the categorical results totals in table 1, the number of cases where the WRF error was statistically significantly lower than the NAM error was 30 out of 48 or 62.5% of the total possible number of cases. The number of cases where the NAM error was significantly lower than the WRF error was 8 out of 48 or 16.7%. The number of cases where the error statistics for the two models were essentially the same was 10 out of 48 or 20.8%.

Another way of interpreting these results is shown in table 2 where the categorical results for each meteorological variable examined are shown for Domain 1 and Domain 2 for the WRF 1-km, WRF 3-km, and NAM 12-km output. As an example, for temperature, the WRF 1-km and WRF 3-km output had better results than the NAM in five of the six cases. The only case where the NAM was better than the WRF was for temperature Bias for Domain 2.

Table 2. Comparison between WRF and NAM for significant differences between the models for Domain 1 and Domain 2.

Variable	**WRF < NAM**	**WRF = NAM**	**NAM < WRF**
2m TMP	5	0	1
2m DP	6	0	0
2m RH	5	0	1
MSLP	1	3	2
10m U-WIND	2	1	3
10m V-WIND	3	3	0
10m WIND	3	3	0
10m RM-WDIR	3	0	1
10m AG-WDIR	2	0	0
All Variables	30	10	8

The percentages in table 2 show that statistically, the WRF outperforms the NAM as evidenced by lower error statistics for the variables of temperature, dew point temperature, relative humidity, row mean wind direction and aggregate wind direction. For wind speed, and V-component wind speed, the degree to which the WRF outperforms the NAM is questionable. For sea level pressure the two models perform roughly the same and for the U-component wind speed, the NAM appears to outperform the WRF by a small margin.

The overall categorical results from table 2 were separated into those which came from model comparisons made in Domain 1 and in Domain 2. Domain 1 comparisons were for the 3-km WRF and the 12-km NAM model and Domain 2 comparisons were for both resolutions of the WRF (1 km and 3 km) and the 12-km NAM model. For Domain 2, the criteria for the categorization allowed for either one of the WRF model’s results to determine which category the comparison produced, this effectively makes the comparison into WRF vs. NAM with no distinction needed as to which WRF (1-km or 3-km) results were used.

Investigating Domain 1 alone (not shown) for all variables except MSLP it appears that the WRF outperforms the NAM, as the WRF has more favorable results in 75% of the cases. For Domain

2 (not shown) the degree to which the WRF outperforms the NAM is reduced for several variables as the result is only 50%.

When comparing the 3-km and 1-km models against the NAM alone, the results are shown in table 3.

Table 3. Categorical percentages for the 1-km and 3-km WRF errors against the 12-km NAM for Domain 2.

Model Resolution	WRF < NAM	WRF = NAM	NAM < WRF
1-km WRF	50	29	21
3-km WRF	42	37	21

Looking at the results of table 3 it can be inferred that there is not much difference when comparing the 1-km and 3-km model against the 12-km NAM; however, we can conclude that the overall data sample would indicate that the WRF performs better than the NAM over the Dugway area for this 31-day study. It should be noted, that only the basic set of weather parameters were tested when considering the surface temperature, moisture, and winds. Analysis of the categorical results suggest that for the majority of meteorological variables the WRF performed statistically better than the NAM. However, the differences in the value of the error statistics between the WRF and the NAM are considered insignificant in terms of real impact of such errors on operations and thus a clear determination of the value added of the WRF over the NAM initialization grid is not possible based on these results alone.

3.4 Comparison of Overall Surface Meteorological Variable Errors for the Three Model Resolutions By Forecast Hour

To better understand model performance it is often interesting to study the hourly results of the models. This may give clues about model strengths, weaknesses, and Biases. Figure 17 shows the surface (2-m) temperature errors at specified forecast lead times for Domain 2. As mentioned previously, the model cycle begins at 0600 UTC (0-h forecast); however, the evaluation period begins 6 h later and runs for 6 h from 1200 UTC to 1800 UTC (6-h to 12-h forecast). Based on the results in figure 17, the best performance of the model temperature forecast was by the 1-km and 3-km WRF at the 9-h forecast valid time (1500 UTC). The forecast error was less than 2 °K, with a near perfect Bias. At 1800 UTC, all three model outputs displayed approximately a 2 °K error with WRF slightly overestimating values and the NAM nearly perfect on the Bias. A larger performance error was noted at 1200 UTC, where all three models reported a 3 °K error and an overforecasting tendency.

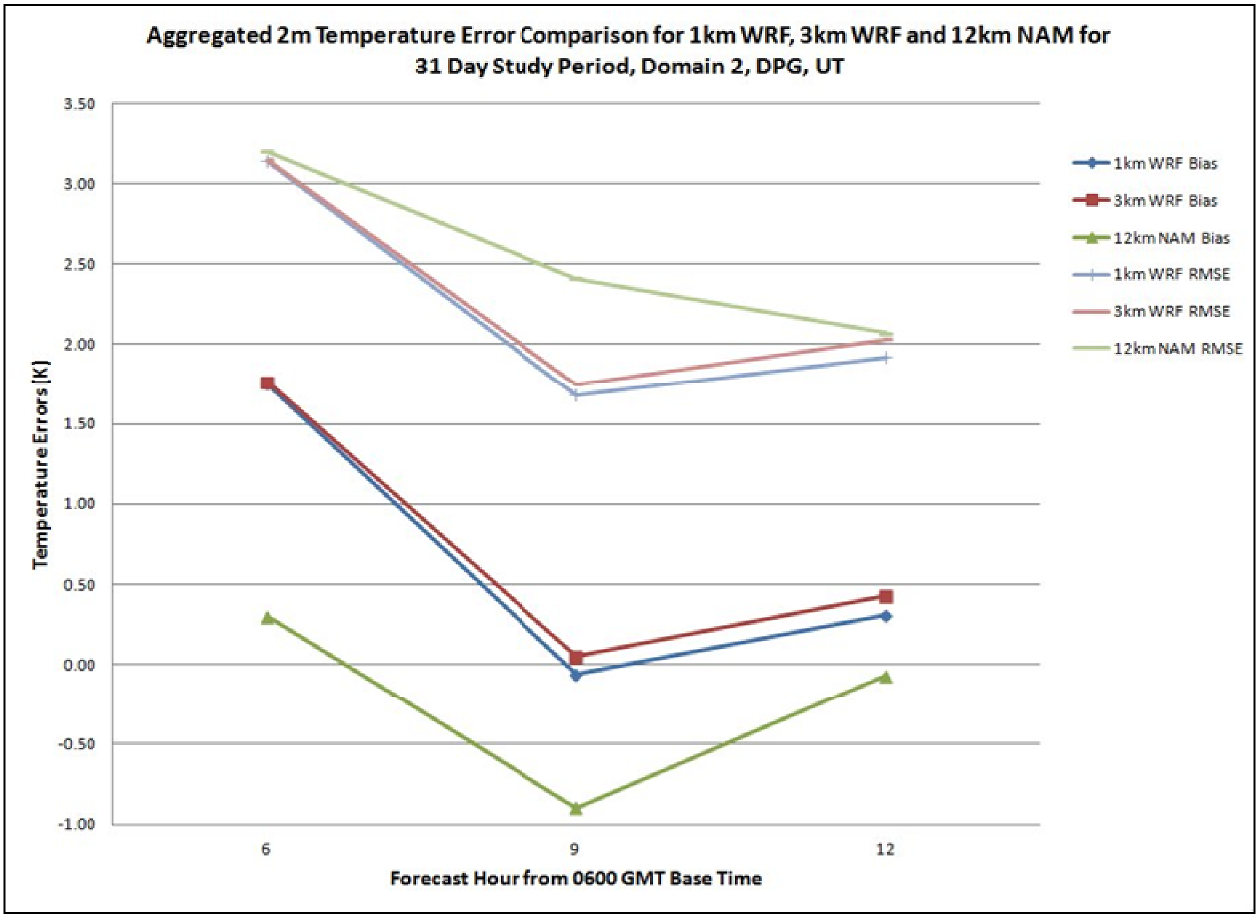

Figure 17. Comparison of the 2-m air temperature errors for the 6-, 9-, and 12-h forecasts valid at 1200, 1500, 1800 UTC for Domain 2.

The surface dew point temperature (2-m) errors at specified forecast times for Domain 2 are displayed in figure 18. The RMSE of the dew point error varies between 2.5 °K and 3.5 °K for the three models, with the NAM displays an overforecasting Bias for all time periods and the WRF underestimating dew point values at 1200 UTC and 1500 UTC, with a very minimal overforecasting Bias at 1800 UTC.

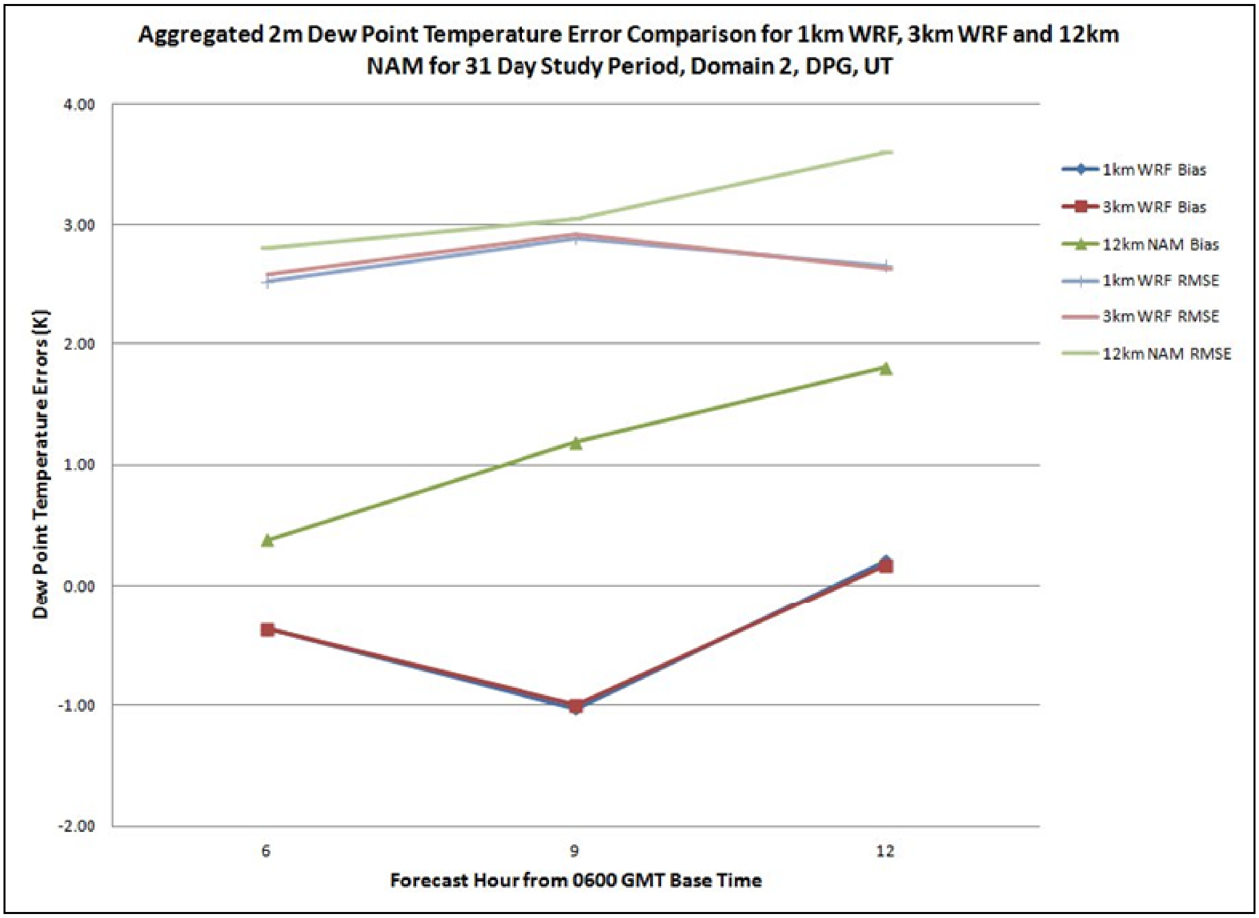

Figure 18. Comparison of the 2-m dew point temperature errors for the 6-, 9-, and 12-h forecasts valid at 1200, 1500, 1800 UTC for Domain 2.

Figure 19 shows the surface wind speed errors at specified forecast times for Domain 2. The spread of the RMSE between the three models was less than 1 m/s, generally between 1.8 m/s and 2.3 m/s; therefore, this error may not be significant. The WRF and NAM generally underforecast the wind speed, although there is a slight overforecasting Bias by the 1-km and 3-km WRF at 1500 UTC. The NAM does show a tendency to underforecast the wind speeds more significantly at 1500 UTC.

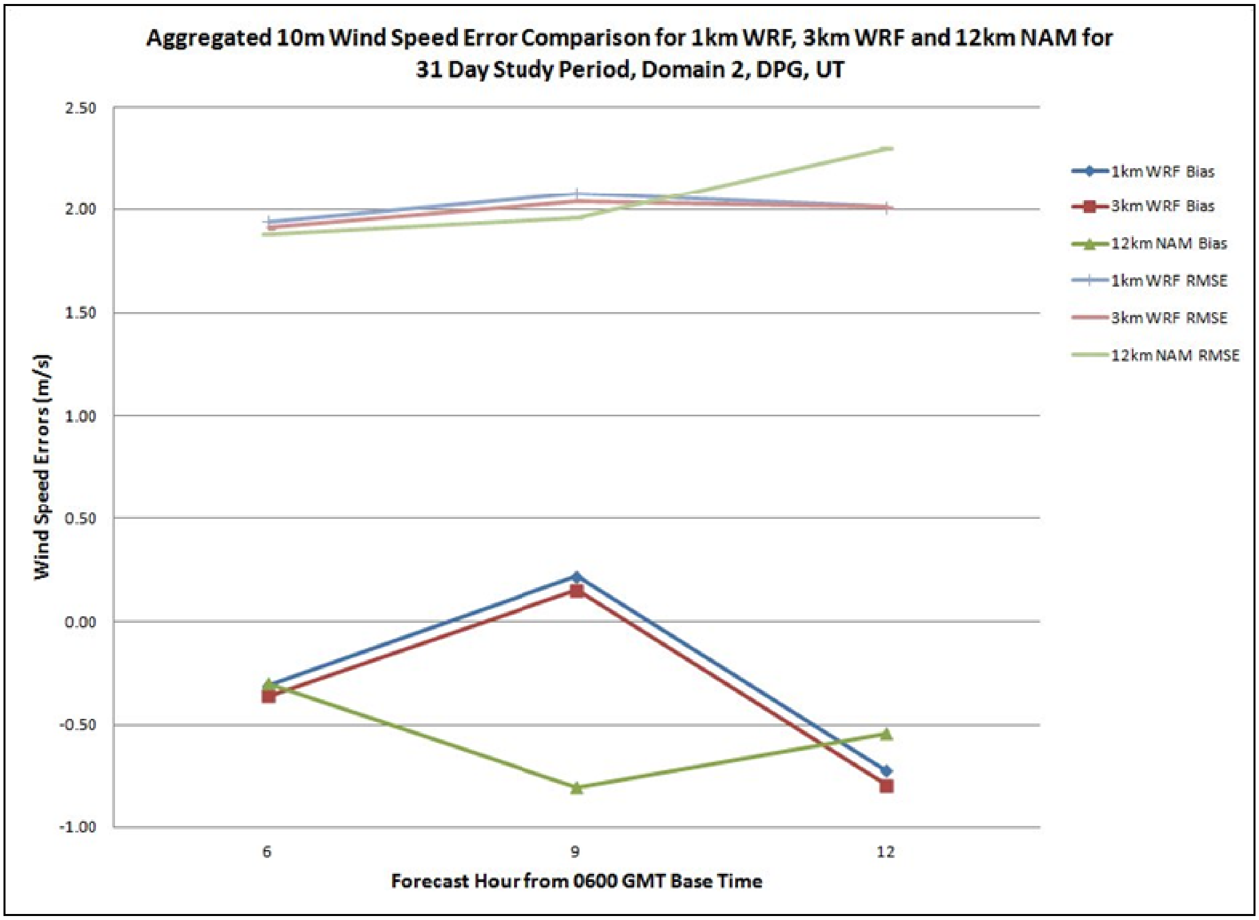

Figure 19. Comparison of the 10-m wind speed errors for the 6-, 9-, and 12-h forecasts valid at 1200, 1500, 1800 UTC for Domain 2.

The surface row mean wind direction errors at specified forecast times for Domain 2 are displayed in figure 20. The calculation of statistics for wind direction are often more difficult due to the variability in wind speeds, particularly during the morning hours in an area of complex terrain, such as Dugway. Setting that concern aside, the marked decrease from 1200 UTC to 1800 UTC in MAE of 25° over 6 h is most likely a significant result and perhaps related to the model adjustments through the morning hours. The model Biases are also of interest as the WRF tends to underforecast the wind direction while the NAM overforecasts the wind directions. These spreads are of interest and shows a significant result that can be important to model users.

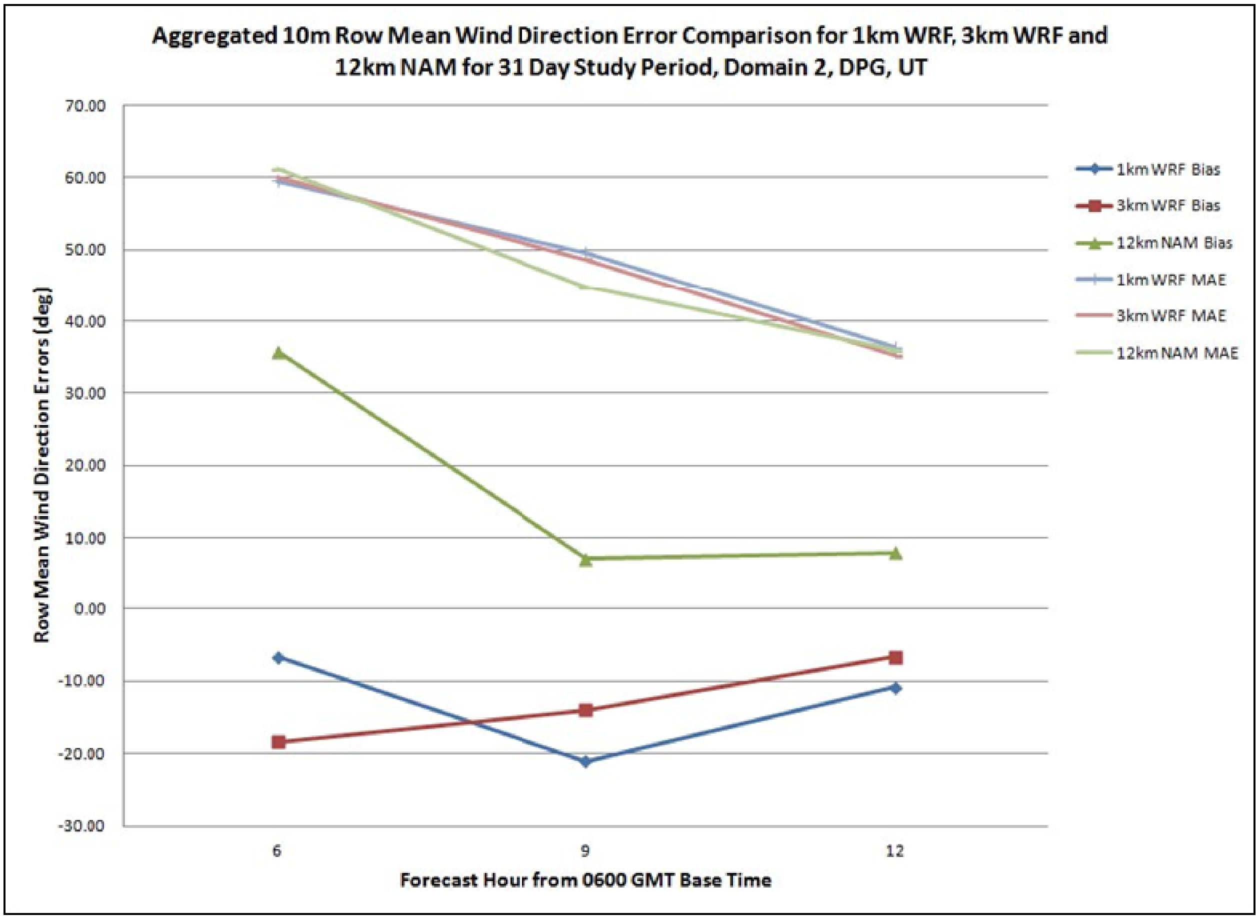

Figure 20. Comparison of the 10-m row mean wind direction errors for the 6-, 9-, and 12-h forecasts valid at 1200, 1500, 1800 UTC for Domain 2.

3.5 Comparison of Overall Upper-Air Meteorological Variable Errors for the Two Model Resolutions

Upper-air meteorological variable error statistics are calculated by MET Point-Stat in much the same way that surface variable statistics are with some notable exceptions. The forecast-observation differences are computed from the model forecasts valid at just the two standard rawinsonde observation times, 0000 UTC and 1200 UTC which are the 18-h and 6-h forecasts respectively and valid at the location of the upper-air stations. For Domain 1 there are two upper-air sounding stations which are Salt Lake City, UT (KSLC) and Elko, NV (KEKO). Domain 2 does not have any regular reporting upper-air stations. Forecast-observation differences are calculated for a range of levels in the vertical and are assigned to the central pressure level value in the middle of each range group. The error statistics are averaged within the following groups in hPa: 225−100, 425−225, 625−425, 775−625, 875−775, 910−875, and 1010−910. Thus, the plots of the upper-air statistics show the error values at discrete levels which are the central value within these groups. The tabular results show both the range of levels for each group, as well as the central value. The error statistics are then aggregated for both upper-air stations for each of the two models which were run in Domain 1, which were the 3-km resolution WRF and the 12-km resolution NAM. This was done separately for the 6-h forecast and the 18-h forecast for both models. Then these results were aggregated over all 31 days of the study period.

The upper-air temperatures errors at 1200 UTC (6-h forecast) for Domain 1 are shown in figure 21. Looking at the RMSE the 12-km NAM shows the error to be slightly less than the 3-km WRF, as it underestimates the forecasted temperatures below 500 hPa. The model Biases diverge near the surface, as the WRF overforecasts temperatures while the NAM underforecasts the temperatures. Above 500 hPa, the close proximity of both model results, coupled with the concurrent overestimation of the forecast, indicates no statistically significant difference between the models. Both models also exhibit the same minimum RMSE (about 1 °K) at mid-atmospheric levels with increasing magnitude above and below this level.

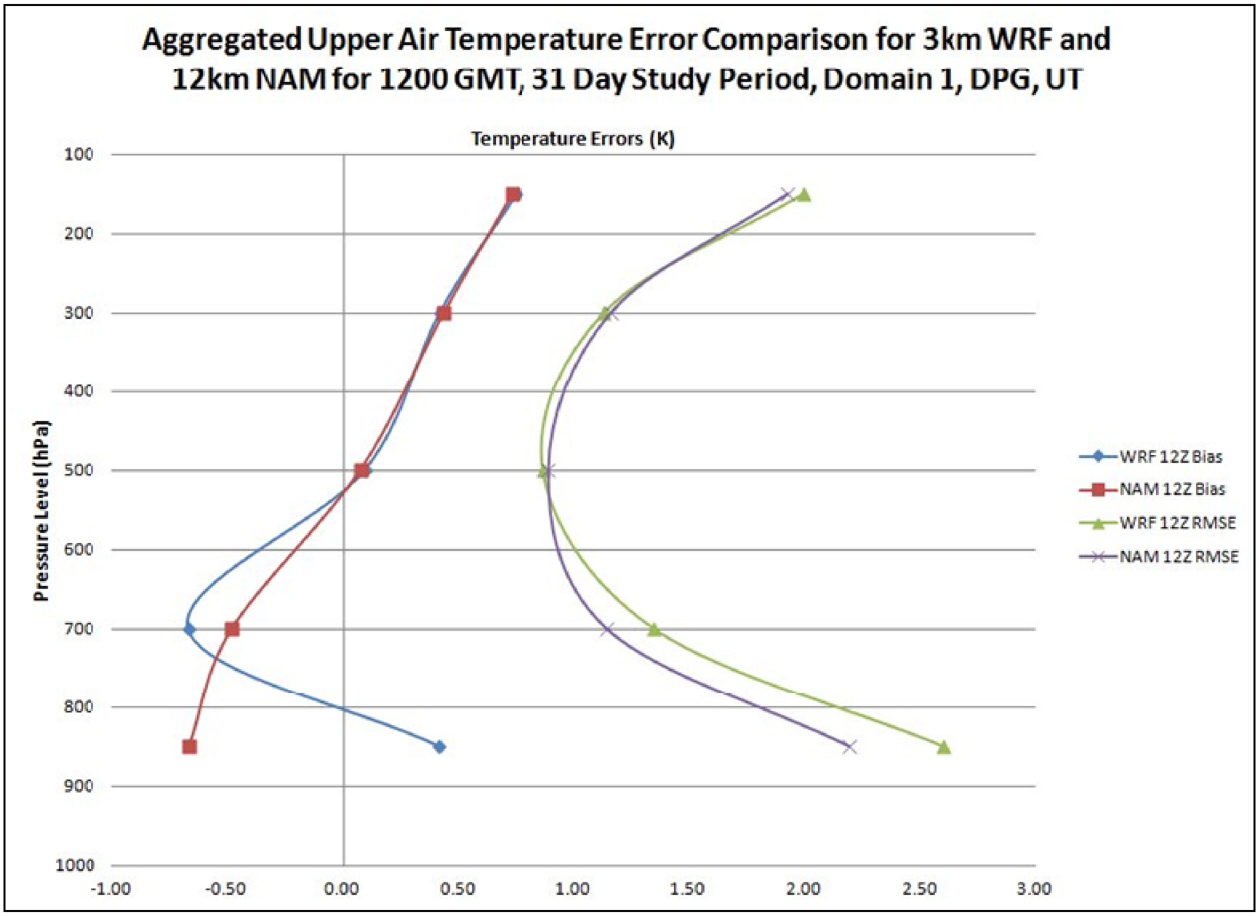

Figure 21. Comparison of the upper-air temperature errors for the 6-h forecast valid at 1200 UTC for Domain 1.

Figure 22 shows the upper-air relative humidity errors at 1200 UTC for Domain 1. Since upper-air forecasts of dew point temperature are not available for the NAM model, the relative humidity is provided instead. The RMSE statistical error between models varies less than 5% with larger errors as the profile increases with height. Both models show a tendency to overforecast the RH, although the WRF 1200 UTC underforecasts the RH at the 850-hPa level. Note that for the WRF the moist Bias tendency above 850 hPa is the opposite tendency from the surface relative humidity forecasts which have a dry Bias.

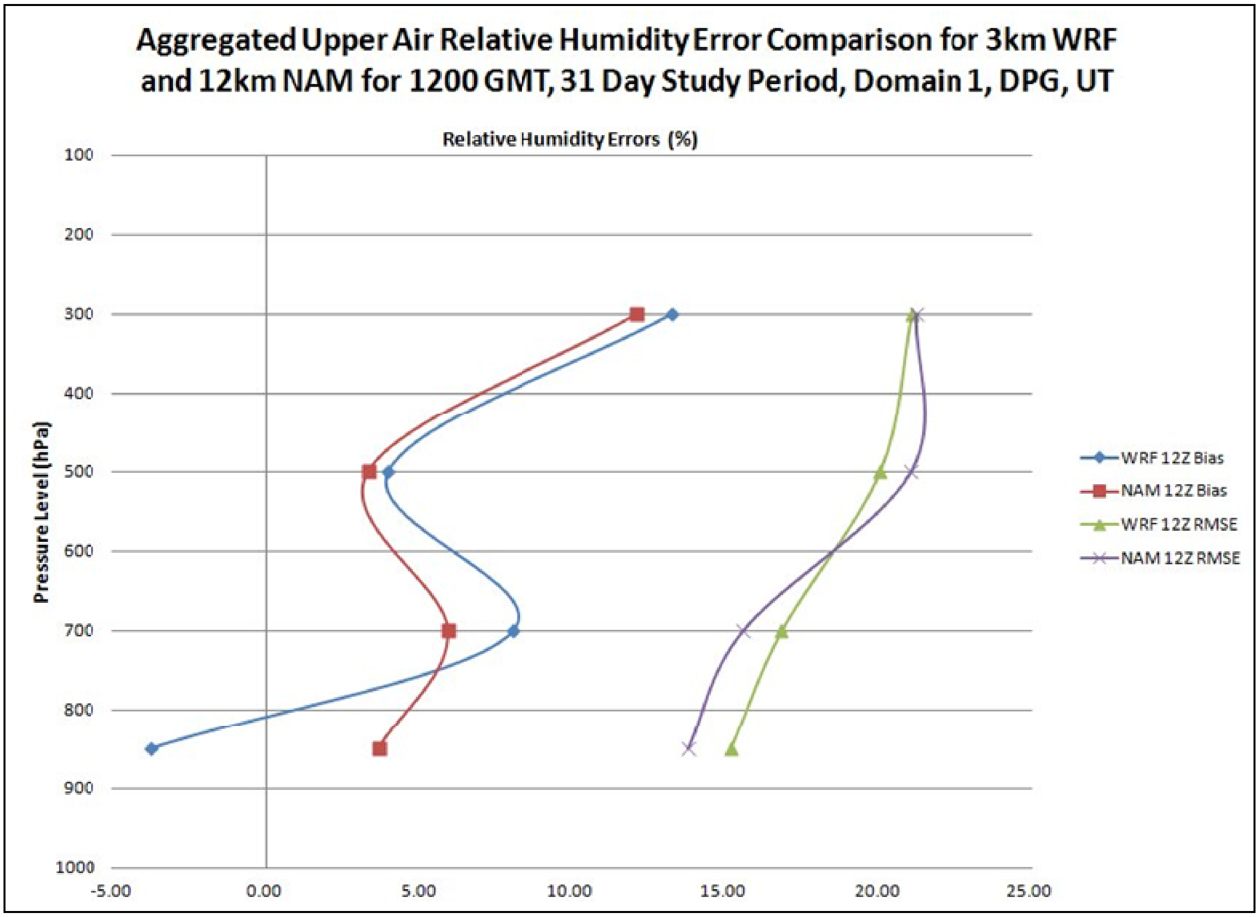

Figure 22. Comparison of the upper-air relative humidity errors for the 6-h forecast valid at 1200 UTC for Domain 1.

Figures 23 and 24 shows the upper-air U-component wind speed errors and V-component wind speed errors respectively at 1200 UTC for Domain 1. It should be noted that the upper-air forecasts of wind speed are not available for the NAM model, so the component wind speed are provided instead. The statistical errors between models again show little variation. The RMSE for both the 3-km WRF and 12-km NAM show similar trends although there is slightly higher RMSE for the V-component than the U-component at 1200 UTC. There is slightly higher Bias in the upper-levels for the U-component than the V-component, although that is not a significant issue. The other noticeable disparity is that there is slight Bias to overforecast the U-component wind speed below 600 hPa and underforecast the V-component winds except below 770 hPa for the 1200 UTC NAM.

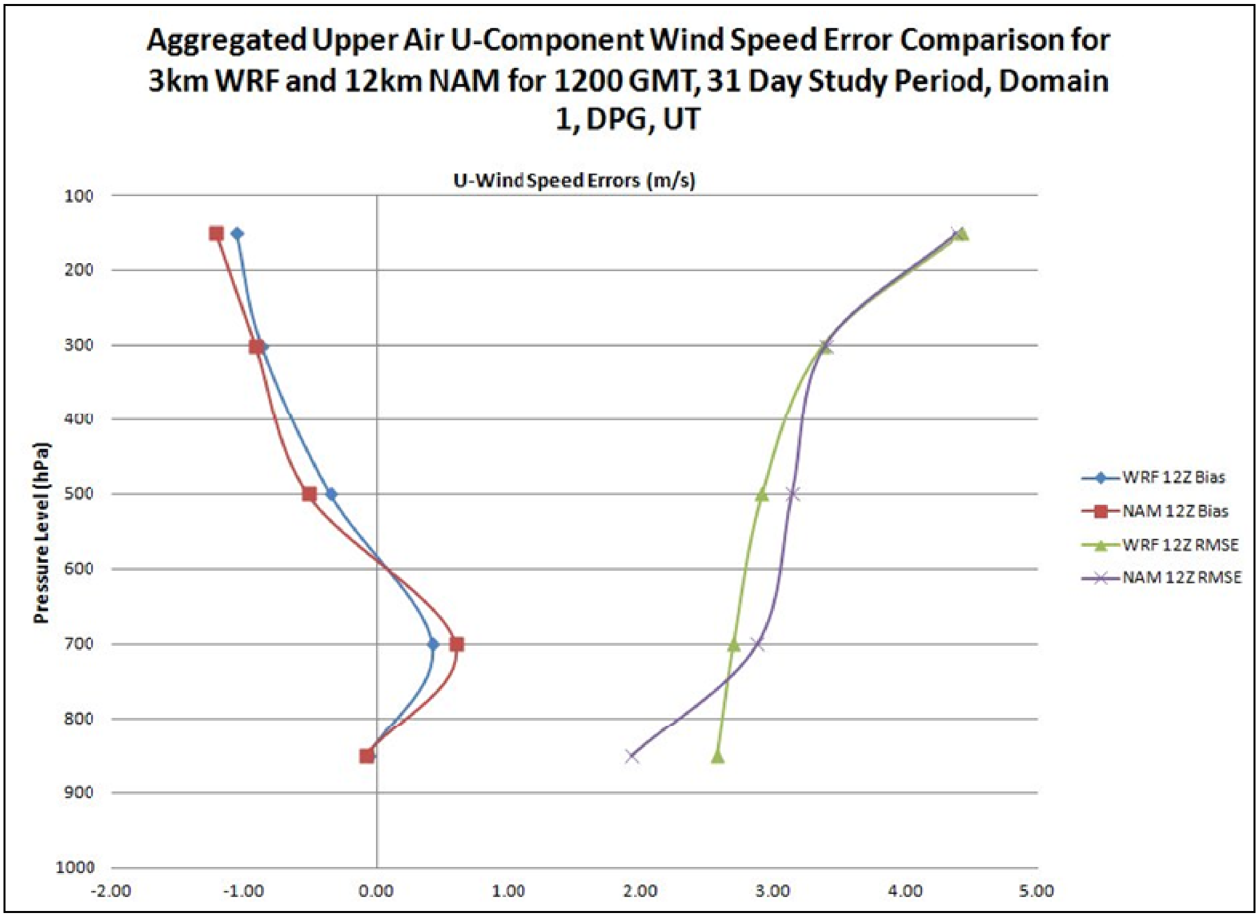

Figure 23. Comparison of the upper-air U-component wind speed errors for the 6-h forecast valid at 1200 UTC for Domain 1.

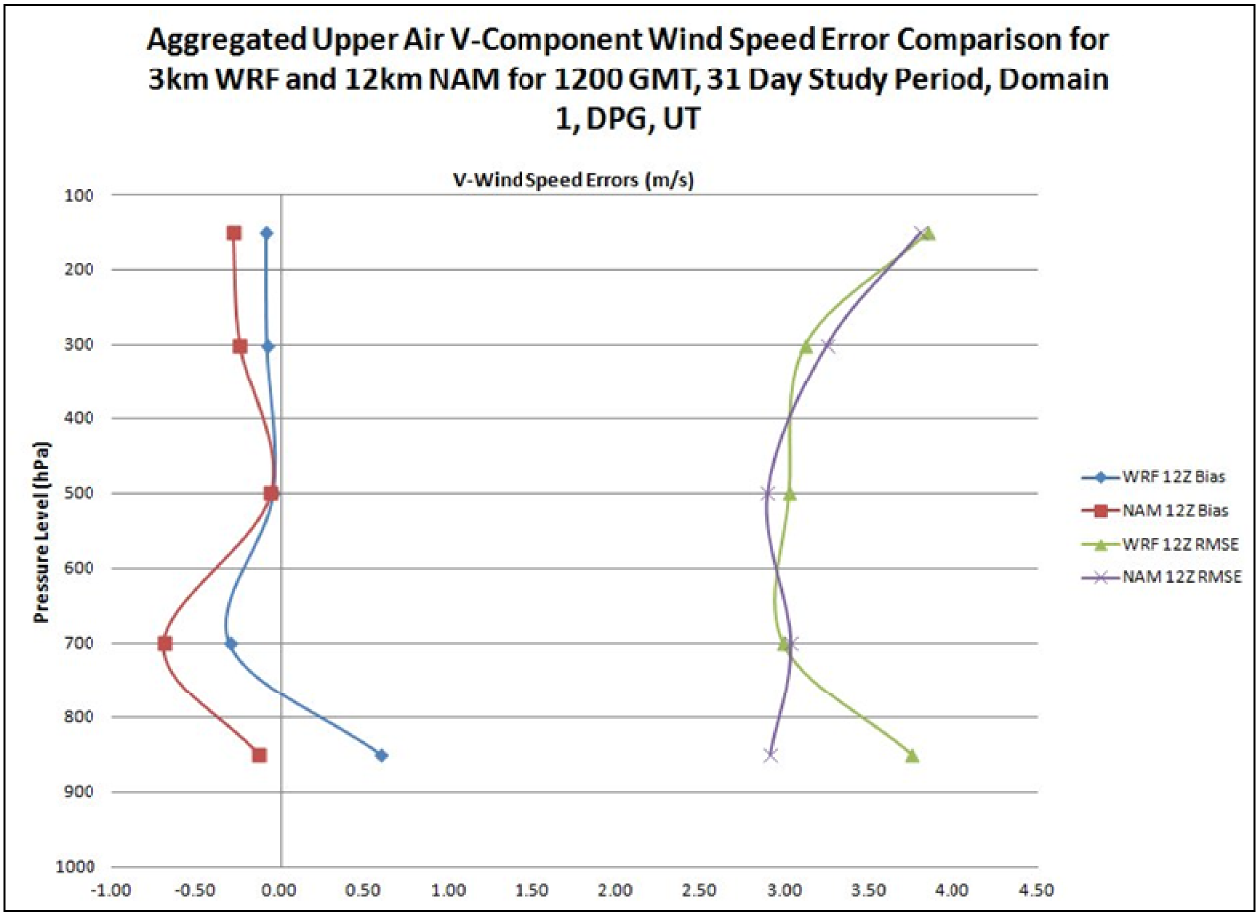

Figure 24. Comparison of the upper-air V-component wind speed errors for the 6-h forecast valid at 1200 UTC for Domain 1.

Figure 25 shows the upper-air row mean wind-direction errors at 1200 UTC for Domain 1. The statistical variation between the two models is not significant. The RMSE were similar from 850 hPa to150 hPa. At lower levels, such as 850 hPa and 700 hPa, the Bias indicates that the NAM underforecasts while the WRF overforecasts wind direction. However, by the 500-hPa level, both models were nearly perfect for their Biases. Additionally, there is the sharp drop in error magnitude (RMSE) from 850 hPA to 500 hPa, by both models. The 5° error for wind direction above 500 hPa is an intriguing statistic; however, it is not surprising given that the wind direction is better-behaved well above the boundary layer and does not vary as much since upper-level winds are dominated by slower-moving dynamic systems while the lower-level winds are influenced by complex interactions between the boundary-layer, surface layer, and complex terrain.

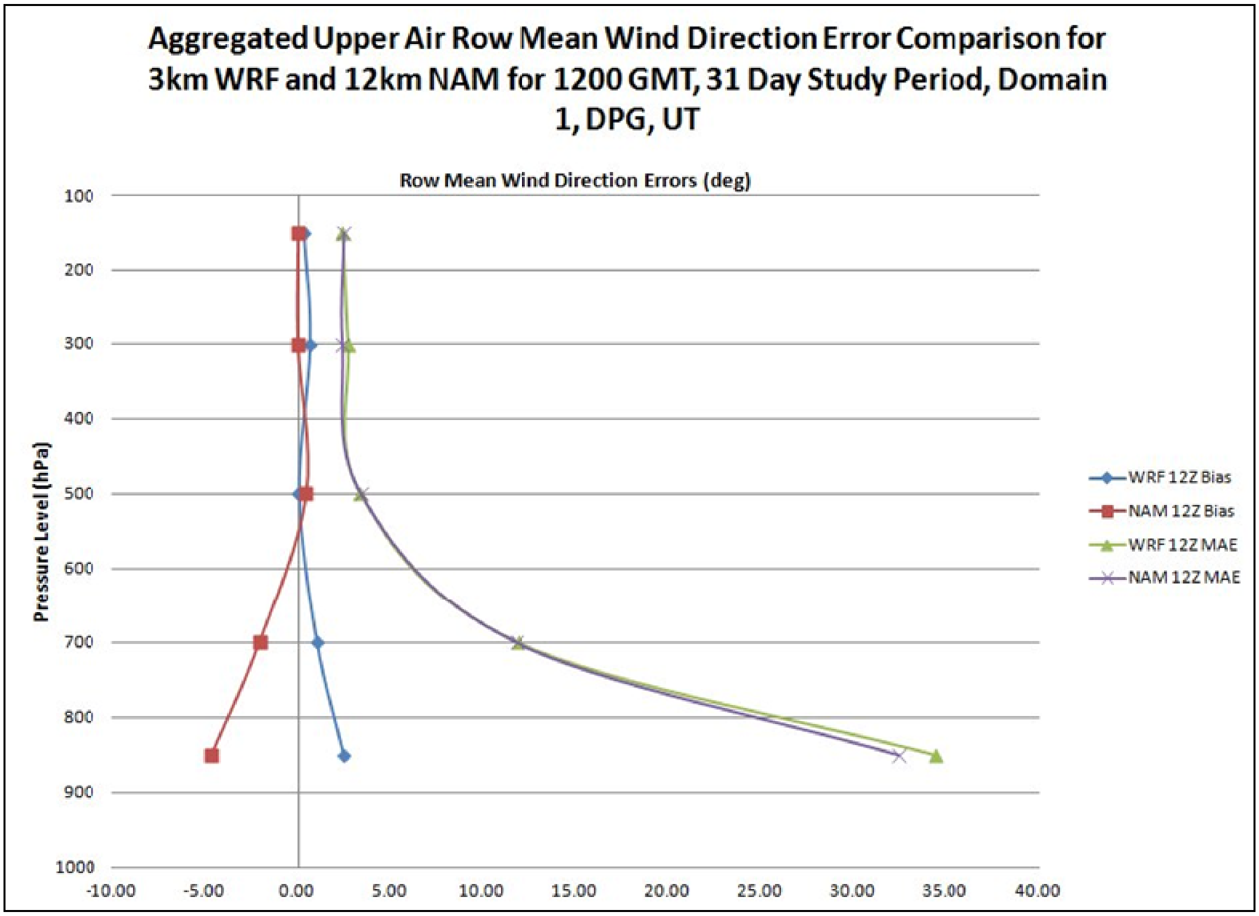

Figure 25. Comparison of the upper-air row mean wind direction errors for the 6-h forecast valid at 1200 UTC for Domain 1.

In general, these results show that there is no significant difference between the NAM and the WRF in terms of upper-air error statistics. Most observed differences in error statistics are found at the lowest levels in or near the boundary layer where model resolution may play a more active role with terrain induced perturbations. While not shown here, but available in the appendix (see figures A-1 to A-5) there is also no significant difference between the 1200 UTC model forecasts and 0000 UTC forecasts.

3.6 Comparison of Results Between a 20-Day Random Study and 31-Day Continuous Study at Dugway

Another comparison was done to study the difference in two WRF studies over the Dugway grid. The first study, conducted in 2010, featured 20 random weather days during the year. The second study was done for 31 consecutive days during May and June of 2011. Table 4 shows the RMSE for the surface variables between the 1-km and 3-km WRF for the two studies over Domain 2. Table 5 displays the Biases for the meteorological parameters for the 1-km and 3-km WRF over Domain 2. While there are some minor difference in skill there are no significant difference seen between the 1-km and 3-km model runs or between the 20-random days or 31-consecutive day study. In reality, the results are almost identical for every variable. Even the higher error in mean sea level pressure is seen in all the cases over the Dugway grid.

Table 4. Comparison of surface RMSE Errors from 2009-10 case studies and from 2011 31-day continuous study.

Variable	**1-km WRF RMSE**	**1-km WRF RMSE**	**3-km WRF RMSE**	**3-km WRF RMSE**
	2009-10 Cases (20 runs)	**2011 Study Period (31 runs)**	**2009-10 Cases (20 runs)**	**2011 Study Period (31 runs)**
TMP (K)	2.54	2.66	2.58	2.70
DPT (K)	2.64	2.89	2.64	2.90
RH (%)	12.41	13.63	12.41	13.70
MSLP	4.49	5.47	4.51	5.58
U (m/s)	2.52	2.42	2.47	2.36
V (m/s)	2.88	2.82	2.84	2.80
Speed (m/s)	2.50	2.46	2.46	2.43
R-M Dir (deg)	NA	NA	NA	NA
AGGR Dir (deg)	NA	NA	NA	NA

Table 5. Comparison of surface Bias errors from 2009-10 case studies and from 2011 31-day continuous study.

Variable	1-km WRF Bias	1-km WRF Bias	3-km WRF Bias	3-km WRF Bias
	2009-10 Cases (20 runs)	2011 Study Period (31 runs)	2009-10 Cases (20 runs)	2011 Study Period (31 runs)
TMP (K)	0.46	0.80	0.49	0.88
DPT (K)	0.29	−0.31	0.28	−0.34
RH (%)	−1.83	−4.69	−1.97	−4.93
MSLP	−1.68	−5.04	−1.78	−5.16
U (m/s)	0.38	0.33	0.35	0.33
V (m/s)	0.13	0.41	0.16	0.41
Speed (m/s)	0.19	−0.22	0.09	−0.27
R-M Dir (deg)	−4.62	−2.49	−3.85	−2.56
AGGR Dir (deg)	54.71	54.15	56.01	55.14

3.7 Discussion of Results

There have been numerous studies of model effectiveness and statistical output. Many of these involve comparisons of different models, case studies of a certain parameter, or models of different resolutions in many locations. This study emphasized a study between 1-km and 3-km WRF models, as well as the 12-km NAM. Analysis of all statistical results showed that even though the differences in the errors of the WRF and NAM were statistically significant in many cases, the magnitude of these errors were not large enough to be considered significant for operational use. The difference between a forecast of 22 °C and 24 °C probably will not have much impact to a Soldier or a mission; however, the difference between −1 °C and +1 °C might be very significant. Part of operational significance does depend on the situation or importance of a weather parameter application in the operation. From a statistical standpoint, that was impossible to quantify. However, what can be concluded from this study is: Showing that the WRF forecasts add significant value over the NAM forecasts is not possible with these statistical results. The unique behavior of the NAM and WRF errors as a function of forecast hour and the fact that both WRF model errors tended to be grouped close together and separated from the NAM error value, suggests that there were inherent differences between the WRF and the NAM. Consequently, nothing conclusive can be said about the value added of one model over the other, based on the results of this study. The upper-air errors for all three models were essentially the same, with minor differences primarily occurring in the lowest layers, in or near the boundary layer, where model resolution may have been a factor.

In a majority of the cases where the overall errors of the NAM and WRF were compared categorically, based on the overlap or non-overlap of the error statistics with confidence intervals, the WRF statistically outperformed the NAM.

The forecast errors of all three models for mean sea level pressure appeared to be excessive, based on the magnitude of the RMSE values, which ranged between 5 hPa and 6 hPa. The sign of the Bias was negative (underforecast). A comparison with a case from AFWA showed the WRF RMSE values for sea level pressure was approximately 2 mb or less with a positive Bias (overforecasting); however, it is uncertain if the error in the current study was due to calculation problems, instrument calibration issues, or is just a natural feature of the model in high-desert and mountainous terrain.

Additionally, there was little difference between the errors of 20 random days and the errors of 31 consecutive days. The RMSE for the 20-day study was almost identical to the RMSE for the 31-day study. While this study was conducted in Utah, it is uncertain if the results would be consistent in other locations.

Perhaps, the most important finding in this study was that model users can have confidence in lower resolution models, such as the 12-km NAM, but there are many advantages to using higher-resolution models. Some of these advantages can be seen in cases where the WRF outperforms the 12-km NAM, but many of them cannot be fully observed with the current available observation network. It may be true that running a model at 1 km cannot be fully appreciated without a dense observation network. While this study did have a reliable and robust source of surface data, the study was still unable to show the value added from using a high-resolution model, such as the 1-km WRF. The advantages may be hidden or still not fully understood in very small-scale wind flows, or a sharp temperature gradient along a sloping terrain, or fog observed in a local valley.

4. Conclusions and Future Work

The ARL has been utilizing the high resolution mesoscale model WRF, for many years. Model validation has always been of interest to the civilian and military forecasting community; however, this validation has proven to be especially difficult when addressing small time- and space-scales, such as those provided by the WRF. ARL has previously used the MET to provide traditional statistical measures. One such study in 2010 used a large number of MADIS observational data. This enhanced set of verification statistics, using grid-to-point evaluations, revealed that the forecasting errors varied depending on the case-study day. It was questioned if this case study approach may have skewed the results towards model performance in these specific types of weather conditions. Additionally, it was found in the 2010 study that the errors for the 1-km and 3-km horizontal resolution WRF versions were statistically the same.

In this current study, these concerns were addressed by running the model over a continuous 31-day period from 27 May 2011 to 26 June 2011. To establish a benchmark comparison for WRF performance, which could later be compared with the WRF-FDDA performance, it was decided to acquire and assess the output of the NAM model. The NAM model was run over the same domains as the WRF model and served as the initialization grid for the WRF, but at a lower horizontal resolution of 12 km. This study emphasized a comparison between 1-km and 3-km WRF models, as well as, the 12-km NAM. Analysis of all the statistical results showed that even though the differences in the WRF and NAM errors were significant in many cases, the magnitude of these errors were not large enough to be considered significant in operations. It was also determined that it was not possible to show that the WRF forecasts added value over the NAM forecasts. Additionally, nothing conclusive can be said about the value added of one model over the other, based on the results of this study. The upper-air errors for all three models were essentially the same, with minor differences primarily occurring in the boundary layer.

The current ARL implementation of the MET and capabilities which combine traditional, fuzzy and spatial techniques is a powerful means for the assessment of the accuracy of the high resolution WRF models. Work to completely automate Point-Stat, Grid-Stat and MODE should be continued.

As an adjunct effort to the implementation of MODE, the integration of a capability to visualize objects in the forecast and observed grids should be accomplished. This approach will provide a means of defining objects which are suitable for use in spatial verification with MODE. Since MODE was developed using discrete objects from precipitation forecasts and observations, the appropriate use of MODE techniques for objects which are derived from fields of continuous meteorological variables such as air temperature, dew point temperature and wind speed by the application of thresholds, needs to be investigated and developed. A visualization capability which has some promise of fulfilling this need is the Integrated Data Viewer (IDV). Potentially, IDV could be used to inspect objects using its tool to apply thresholds to decide which objects best fit the criteria required by MODE before running MODE.

The potential outcome of using MODE for spatial verification is to assess the performance of decision aids used by the Army. These tools use thresholds derived from system and mission rules which relate their success or failure to whether the meteorological criteria are met or not met. Thus, if a forecast area of high temperatures will adversely impact an operation, as predicted by the WRF model, the accuracy of that area or “object” needs to be assessed.

5. References

Air Force Weather Agency. 16th Weather Squadron/WXN Meteorological Models. https://weather.afwa.af.mil/host_home/DNXM/ (accessed December 2011).

Cornell University Statistical Consulting Unit, StatNews #73. http://www.cscu.cornell.edu/news/statnews/stnews73.pdf (accessed February 2012).

Huschke, R. *Glossary of Meteorology*; American Meteorological Society: Boston, MA, 1959.

MADIS Meteorological Surface Quality Control. http://www-sdd.fsl.noaa.gov/MADIS/madis_sfc_qc.html (accessed February 2012).

National Center for Atmospheric Research. *Model Evaluation Tools Version 2.0 User's Guide*; Developmental Testbed Center: Boulder, CO, 2009.

Orlanski, I. Rational Subdivision of Scales for Atmospheric Processes. *Bull. Amer. Meteor. Soc*. **1975**, *56*, 527–530.

Raby et al. *Traditional Statistical Measures Comparing Weather Research and Forecast Model Output to Observations Centered Over Utah*; ARL-TR-5422; U.S. Army Research Laboratory: White Sands Missile Range, NM, 2011.

Sauter et al. *Traditional Statistical Measures Comparing Weather Research and Forecast Model Output to Observations Centered Over Utah*; ARL internal report; U.S. Army Research Laboratory: White Sands Missile Range, NM, 2009.

Wikipedia. Mesoscale Meteorology. http://en.wikipedia.org/wiki/Mesoscale_meteorology (accessed January 2012).

INTENTIONALLY LEFT BLANK.

Appendix. Tabular and Additional Plotted Error Statistics for Surface and Upper-Air Meteorological Variables for the Three Models

This appendix contains tables and additional graphs of the error statistics of Bias or Mean Error (ME), Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and the total number of matched forecast-observation pairs (TOTAL) used in calculating the statistics for the following surface and upper-air meteorological variables:

- Air temperature (degrees Kelvin)
- Dew point temperature (degrees Kelvin)
- Relative humidity (percent)
- Mean sea level pressure (HectoPascals, 0-m level)
- U-component wind speed (meters/second)
- V-component wind speed (meters/second)
- Wind speed (meters/second)
- Row mean wind direction (degrees)
- Aggregate wind direction (degrees)

Note: MET does not calculate RMSE for wind direction. MET does not calculate MAE for aggregate wind direction.

The following page lists the order of the figures and tables in the appendix and each reference is linked to the corresponding figure or table.

List of figures and tables for the appendix:

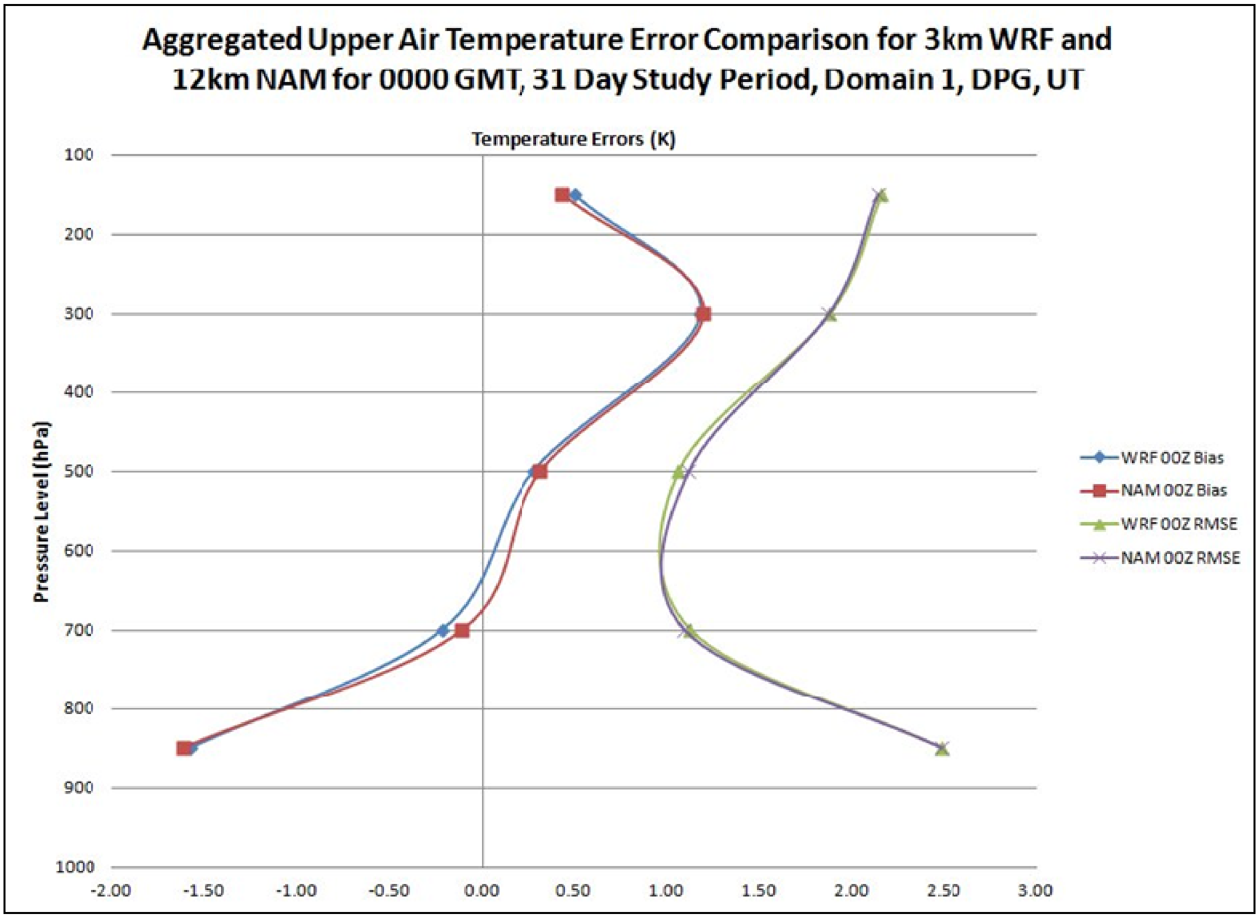

Figure A-1. Comparison of the upper-air temperature errors for the 18-h forecast valid at 0000 UTC, Domain 1.

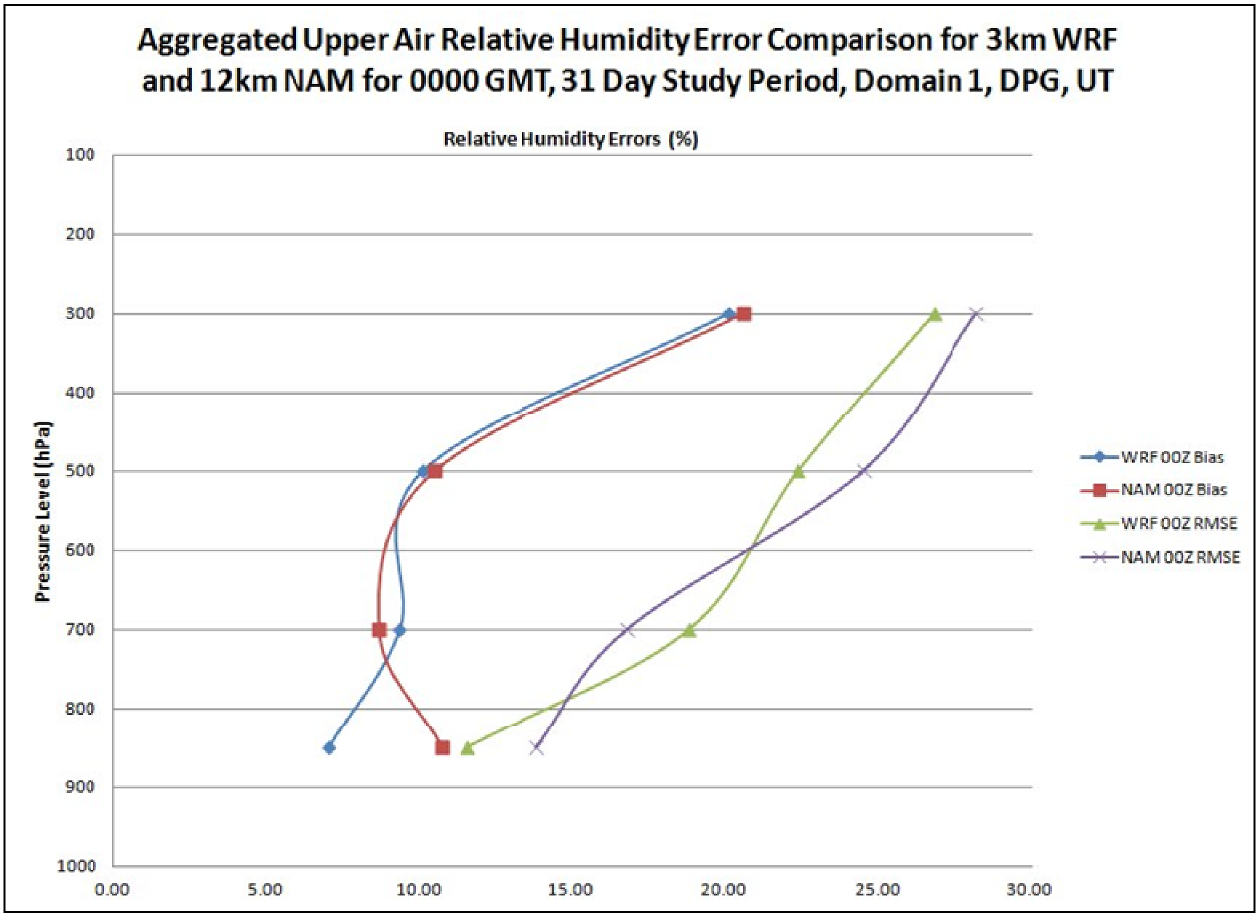

Figure A-2. Comparison of the upper-air relative humidity errors for the 18-h forecast valid at 0000 UTC, Domain 1.

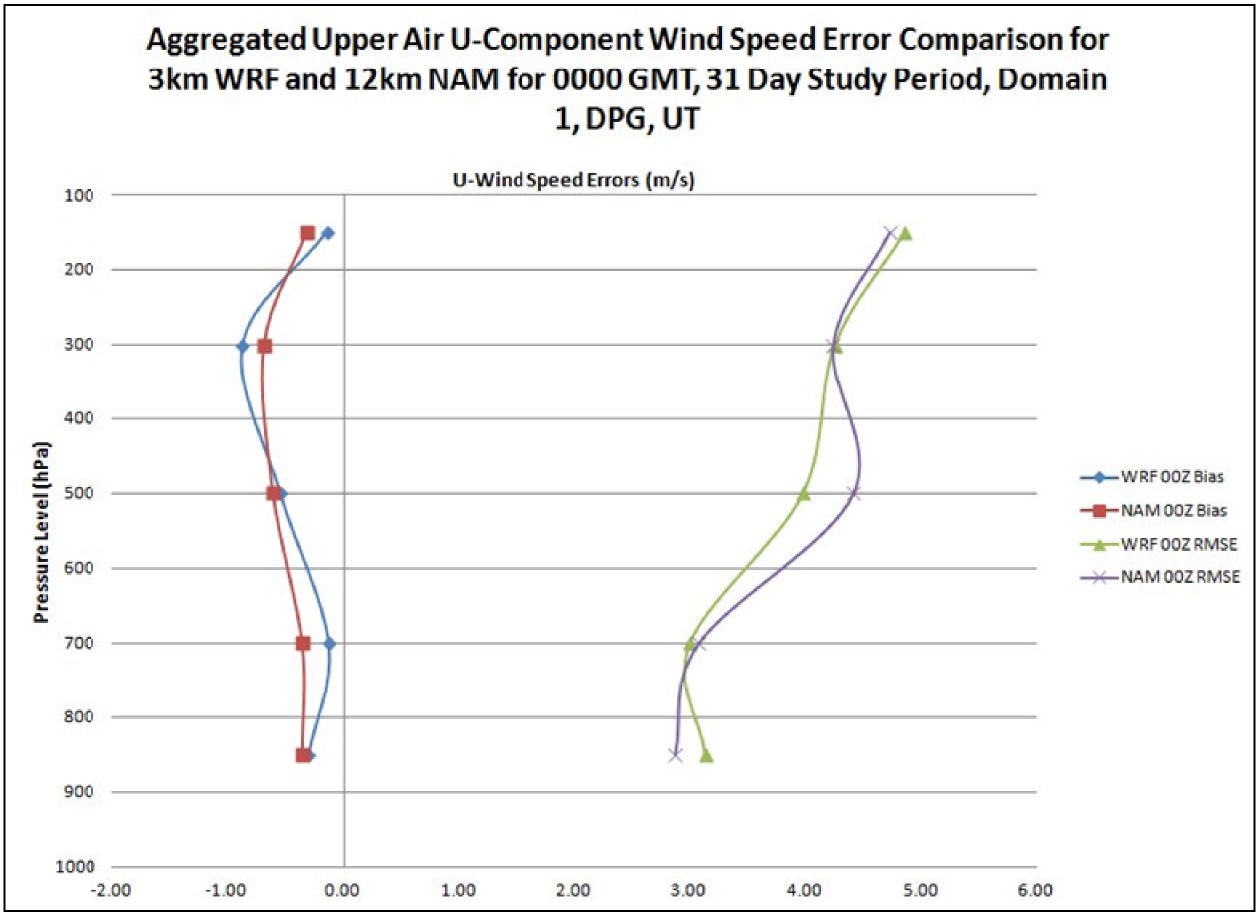

Figure A-3. Comparison of the upper-air U-component wind speed errors for the 18-h forecast valid at 0000 UTC, Domain 1.

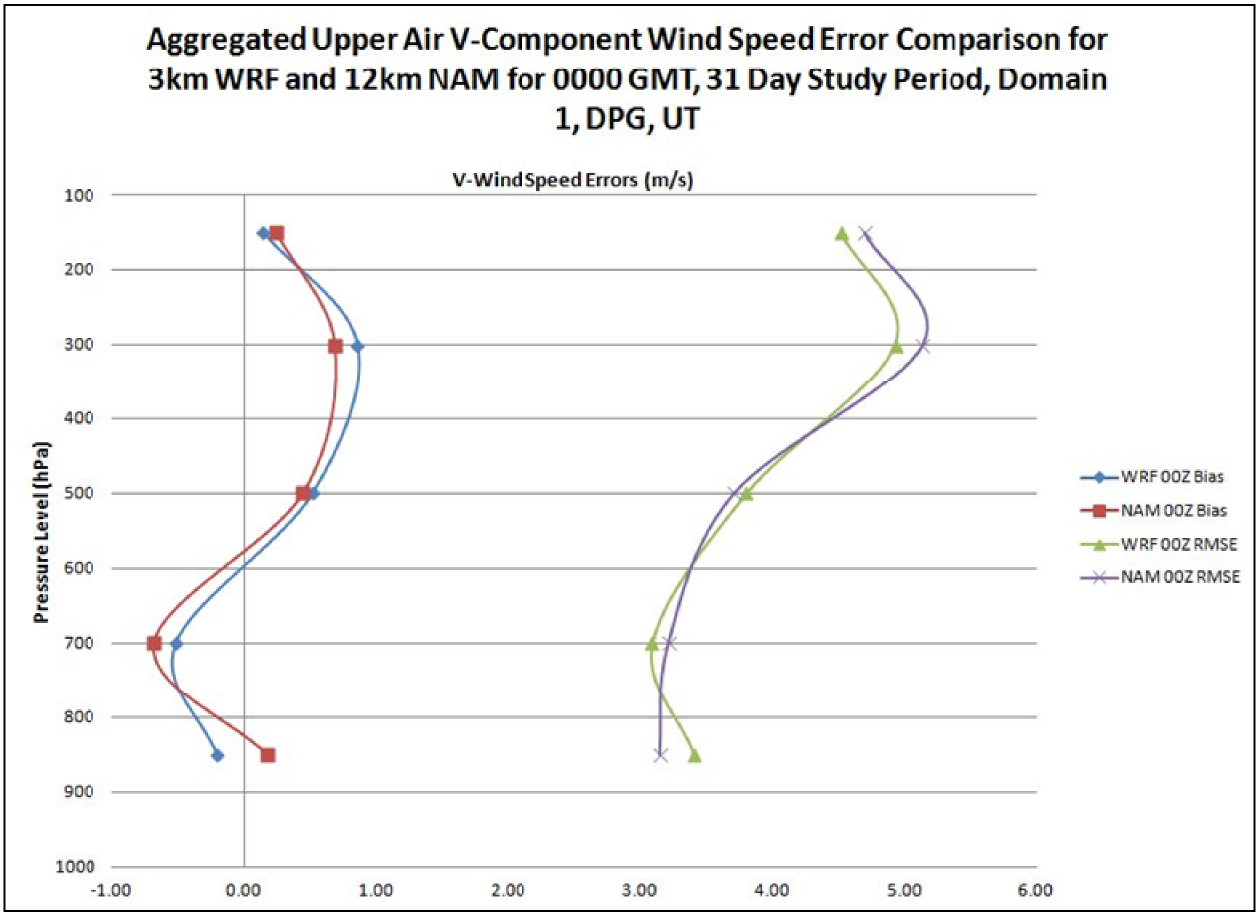

Figure A-4. Comparison of the upper-air V-component wind speed errors for the 18-h forecast valid at 0000 UTC, Domain 1.

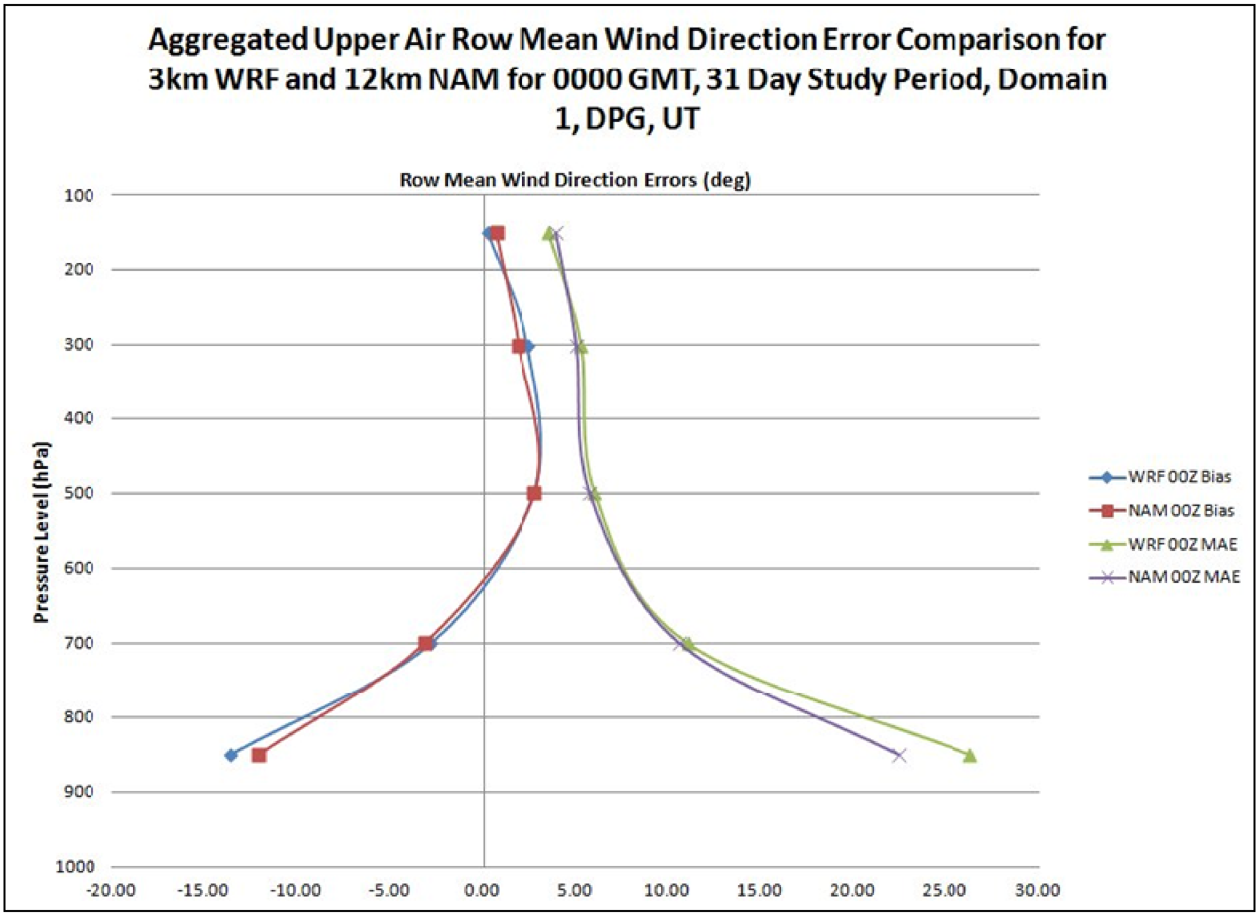

Figure A-5. Comparison of the upper -air row mean wind direction errors for the 18-h forecast valid at 0000 UTC, Domain 1.

Table A-1. Overall Bias statistics for surface meteorological variables aggregated over the 31-day study period for Domain 2.

VRBL	1-km WRF Bias (# F-O pairs)	3-kmWRF Bias (# F-O pairs)	12-km NAM Bias (# F-O Pairs)
TMP (K)	0.80 (5096)	0.88 (5096)	−0.07 (4865)
DPT (K)	−0.31 (5128)	−0.34 (5128)	1.18 (4869)
RH (%)	−4.69 (5129)	−4.93 (5129)	1.53 (4870)
MSLP (hPa)	−5.04 (3221)	−5.16 (3221)	−5.07 (3221)
U (m/s)	0.33 (5107)	0.33 (5107)	0.30 (4844)
V (m/s)	0.41 (5107)	0.41 (5107)	0.73 (4844)
Speed (m/s)	−0.22 (5149)	−0.27 (5149)	−0.58 (4886)
R-M Dir (deg)	−2.49 (278)	−2.56 (278)	14.76 (277)
AGGR Dir (deg)	54.15 (4645)	55.14 (4645)	79.04 (4398)

Table A-2. Overall MAE statistics for surface meteorological variables aggregated over the 31-day study period for Domain 2.

VRBL	1-km WRF MAE	3-km WRF MAE	12-km NAM MAE
TMP (K)	1.99	2.03	2.14
DPT (K)	2.24	2.26	2.71
RH (%)	9.86	9.90	11.07
MSLP	5.08	5.19	5.14
U (m/s)	1.81	1.78	1.68
V (m/s)	2.14	2.14	2.15
Speed (m/s)	1.85	1.82	1.75
R-M Dir (deg)	45.20	55.14	43.84
AGGR Dir (deg)	NA	NA	NA

Table A-3. Overall RMSE statistics for surface meteorological variables aggregated over the 31-day study period for Domain 2.

VRBL	1-km WRF RMSE	3-km WRF RMSE	12-km NAM RMSE
TMP (K)	2.66	2.70	2.82
DPT (K)	2.89	2.90	3.49
RH (%)	13.63	13.70	14.50
MSLP	5.47	5.58	5.62
U (m/s)	2.42	2.36	2.19
V (m/s)	2.82	2.80	2.81
Speed (m/s)	2.46	2.43	2.37
R-M Dir (deg)	NA	NA	NA
AGGR Dir (deg)	NA	NA	NA

Table A-4. Overall Bias statistics for surface meteorological variables aggregated over the 31-day study period for Domain 1.

VRBL	1-km WRF Bias (# F-O pairs)	3-kmWRF Bias (# F-O pairs)	12-km NAM Bias (# F-O pairs)
TMP (K)	NA	−0.20 (143540)	−0.23 (143965)
DPT (K)	NA	−0.40 (102212)	0.99 (102646)
RH (%)	NA	−2.21 (102642)	3.45 (103076)
MSLP	NA	−3.63 (23006)	−3.07 (23040)
U (m/s)	NA	0.41 (104794)	0.24 (105030)
V (m/s)	NA	0.31 (104794)	0.42 (105030)
Speed (m/s)	NA	0.47 (106356)	0.67 (106592)
R-M Dir (deg)	NA	−1.59 (279)	6.05 (279)
AGGR Dir (deg)	NA	7.28 (77647)	13.56 (77778)

Table A-5. Overall MAE statistics for surface meteorological variables aggregated over the 31-day study period for Domain 1.

VRBL	1-km WRF MAE	3-km WRF MAE	12-km NAM MAE
TMP (K)	NA	2.06	2.56
DPT (K)	NA	2.46	2.67
RH (%)	NA	10.26	11.68
MSLP	NA	3.98	3.62
U (m/s)	NA	1.79	1.99
V (m/s)	NA	1.92	1.97
Speed (m/s)	NA	1.80	1.86
R-M Dir (deg)	NA	22.78	25.09
AGGR Dir (deg)	NA	NA	NA

Table A-6. Overall RMSE statistics for surface meteorological variables aggregated over the 31-day study period for Domain 1.

VRBL	1-km WRF RMSE	3-km WRF RMSE	12-km NAM RMSE
TMP (K)	NA	2.65	3.25
DPT (K)	NA	3.21	3.50
RH (%)	NA	13.85	15.07
MSLP	NA	6.34	6.15
U (m/s)	NA	2.43	2.53
V (m/s)	NA	2.62	2.62
Speed (m/s)	NA	2.41	2.40
R-M Dir (deg)	NA	NA	NA
AGGR Dir (deg)	NA	NA	NA

Table A-7. Error statistics for surface meteorological variables aggregated over the 31-day study period for the 6-,9- and 12-h forecasts valid at 1200, 1500, and 1800 UTC, Domain 2.

	2-m Temperature				**2-m Dew Point**				
Hour	**Total**	**ME**	**MAE**	**RMSE**	**Total**	**ME**	**MAE**	**RMSE**	
1-km WRF									
6	509	1.75	2.40	3.14	544	−0.37	1.95	2.53	
9	587	−0.06	1.34	1.68	581	−1.03	2.22	2.89	
12	627	0.31	1.49	1.92	624	0.21	2.04	2.66	
3-km WRF									
6	509	1.76	2.40	3.15	544	−0.37	1.99	2.59	
9	587	0.05	1.41	1.75	581	−1.00	2.24	2.92	
12	627	0.43	1.57	2.03	624	0.17	2.04	2.64	
12-km NAM									
6	491	0.30	2.52	3.20	516	0.38	2.10	2.81	
9	556	−0.90	1.83	2.41	551	1.19	2.34	3.05	
12	596	−0.07	1.52	2.07	593	1.81	2.70	3.60	
					10-m Wind Direction				
	10-m Wind Speed				**Row Mean**				**Aggr**
Hour	**Total**	**ME**	**MAE**	**RMSE**	**Total**	**ME**	**MAE**	**Total**	**ME**
1-km WRF									
6	549	−0.31	1.48	1.94	30	−6.61	59.49	437	−168.31
9	581	0.22	1.58	2.08	31	−21.09	49.57	472	98.46
12	628	−0.72	1.50	2.01	31	−10.86	36.31	587	21.93
3-km WRF									
6	549	−0.36	1.44	1.91	30	−18.37	59.96	437	−171.98
9	581	0.15	1.53	2.04	31	−13.96	48.67	472	106.34
12	628	−0.79	1.50	2.01	31	−6.63	35.21	587	20.87
12-km NAM									
6	520	−0.30	1.40	1.88	30	35.77	61.18	409	−124.02
9	550	−0.80	1.46	1.96	31	7.14	44.69	451	113.09
12	597	−0.54	1.68	2.30	31	7.98	35.97	557	17.18

Table A-8. Upper-air error statistics for meteorological variables for 3-km WRF forecast valid 1200 UTC, Domain 1.

WRF/12Z											
Pressure (hPa)		**Temperature (K)**				**Dew Point (K)**					
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**		
150	225−100	1096	0.75	1.60	1.99	NA	NA	NA	NA		
300	425−225	656	0.42	0.83	1.13	366	3.98	5.17	6.67		
500	625−425	684	0.09	0.70	0.87	677	2.55	5.32	7.67		
700	775−625	316	−0.67	0.99	1.35	314	2.26	3.58	5.48		
850	875−775	301	0.41	1.93	2.59	301	−0.80	2.67	3.65		
900	910−875	1	−0.83	0.83	0.83	1	−1.43	1.43	1.43		
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA		
WRF/12Z											
Pressure (hPa)		**Rel Humidity (%)**				**Height (m)**					
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**		
150	225−100	NA	NA	NA	NA	515	7.06	11.67	15.11		
300	425−225	366	13.24	17.35	21.03	418	−0.36	7.52	9.79		
500	625−425	677	3.98	15.57	20.00	348	−3.92	6.59	7.95		
700	775−625	314	8.04	12.58	16.81	296	−2.42	5.53	6.95		
850	875−775	301	−3.73	11.72	15.14	243	−0.20	5.78	7.74		
900	910−875	1	−1.96	1.96	1.96	1	−3.16	3.16	3.16		
1000	1010−910	NA	NA	NA	NA	121	11.40	15.13	19.71		
WRF/12Z											
Pressure (hPa)		**U-comp (m/s)**				**V-comp (m/s)**					
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**		
150	225−100	605	−1.06	3.64	4.43	605	−0.09	3.04	3.84		
300	425−225	452	−0.87	2.63	3.36	452	−0.08	2.39	3.11		
500	625−425	348	−0.35	2.19	2.91	348	−0.04	2.38	3.02		
700	775−625	296	0.41	2.06	2.69	296	−0.30	2.19	2.99		
850	875−775	212	−0.07	1.90	2.56	212	0.59	2.69	3.74		
900	910−875	1	1.17	1.17	1.17	1	−1.95	1.95	1.95		
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA		
WRF/12Z											
						Wind Direction (deg)					
Pressure (hPa)		**Wind Speed (m/s)**				**ROW_MEAN**				**AGGR**	
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**TOTAL**	**ME**	**MAE**
150	225−100	605	−1.25	3.78	4.61	31	0.27	2.36	604	0.60	NA
300	425−225	452	−1.40	2.69	3.43	31	0.60	2.64	452	0.50	NA
500	625−425	348	−0.31	2.38	3.11	31	0.01	3.31	348	0.35	NA
700	775−625	296	−0.20	2.02	2.89	31	0.98	11.85	290	−4.09	NA
850	875−775	212	1.07	2.40	3.35	31	2.44	34.34	185	8.94	NA
900	910−875	1	−0.64	0.64	0.64	1	−172.81	172.81	1	−172.81	NA
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table A-9. Upper-air error statistics for meteorological variables for 12-km NAM forecast valid 1200 UTC, Domain 1.

NAM/12Z										
Pressure (hPa)		**Temperature (K)**				**Dew Point (K)**				
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**	
150	225−100	1096	0.73	1.56	1.92	NA	NA	NA	NA	
300	425−225	656	0.43	0.83	1.16	NA	NA	NA	NA	
500	625−425	684	0.07	0.71	0.89	NA	NA	NA	NA	
700	775−625	316	−0.48	0.86	1.14	NA	NA	NA	NA	
850	875−775	301	−0.67	1.73	2.19	NA	NA	NA	NA	
900	910−875	1	−0.36	0.36	0.36	NA	NA	NA	NA	
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA	

NAM/12Z										
Pressure (hPa)		**Rel Humidity (%)**				**Height (m)**				
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**	
150	225−100					515	9.85	13.93	17.59	
300	425−225	366	12.11	17.65	21.18	418	2.12	7.92	10.30	
500	625−425	677	3.32	16.21	21.02	348	−0.26	6.53	7.89	
700	775−625	314	5.92	11.53	15.56	296	0.40	5.88	7.66	
850	875−775	301	3.65	10.93	13.76	243	1.61	5.97	8.58	
900	910−875	1	−7.81	7.81	7.81	1	1.38	1.38	1.38	
1000	1010−910	NA	NA	NA	NA	121	19.62	20.37	24.63	

NAM/12Z										
Pressure (hPa)		**U-comp (m/s)**				**V-comp (m/s)**				
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**	
150	225−100	605	−1.21	3.57	4.38	605	−0.28	2.98	3.79	
300	425−225	452	−0.92	2.65	3.39	452	−0.24	2.46	3.24	
500	625−425	348	−0.53	2.35	3.14	348	−0.06	2.33	2.89	
700	775−625	296	0.59	2.29	2.87	296	−0.70	2.22	3.03	
850	875−775	212	−0.09	1.48	1.93	212	−0.13	2.24	2.91	
900	910−875	1	0.81	0.81	0.81	1	−0.37	0.37	0.37	
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA	

NAM/12Z											
						Wind Direction (deg)					
Pressure (hPa)		**Wind Speed (m/s)**				**ROW_MEAN**				**AGGR**	
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**TOTAL**	**ME**	**MAE**
150	225−100	NA	NA	NA	NA	31	−0.02	2.46	604	0.23	NA
300	425−225	NA	NA	NA	NA	31	−0.03	2.37	452	0.17	NA
500	625−425	NA	NA	NA	NA	31	0.42	3.37	348	0.54	NA
700	775−625	NA	NA	NA	NA	31	−2.10	11.76	290	−8.10	NA
850	875−775	NA	NA	NA	NA	31	−4.65	32.32	185	1.46	NA
900	910−875	NA	NA	NA	NA	1	−35.08	35.08	1	−35.08	NA
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table A-10. Upper-air error statistics for meteorological variables for 3-km WRF forecast valid 0000 UTC, Domain 1.

WRF/00Z										
Pressure (hPa)		**Temperature (K)**				**Dew Point (K)**				
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**	
150	225−100	1074	0.50	1.65	2.15	NA	NA	NA	NA	
300	425−225	582	1.18	1.39	1.88	274	6.52	7.38	8.88	
500	625−425	588	0.28	0.79	1.06	575	4.13	6.26	8.60	
700	775−625	270	−0.22	0.85	1.12	270	2.22	4.40	6.16	
850	875−775	202	−1.58	1.98	2.48	202	2.04	3.05	3.68	
900	910−875	1	−4.06	4.06	4.06	1	−0.16	0.16	0.16	
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA	

WRF/00Z										
Pressure (hPa)		**Rel Humidity (%)**				**Height (m)**				
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**	
150	225−100	NA	NA	NA	NA	514	22.15	22.46	25.96	
300	425−225	274	20.07	22.88	26.78	419	6.39	11.07	15.98	
500	625−425	575	10.05	17.54	22.34	333	−0.78	5.93	7.81	
700	775−625	270	9.30	14.71	18.78	273	−2.43	5.13	6.68	
850	875−775	202	6.97	9.07	11.48	234	−2.74	6.42	8.10	
900	910−875	1	7.32	7.32	7.32	1	−4.91	4.91	4.91	
1000	1010−910	NA	NA	NA	NA	118	−6.91	10.42	13.42	

WRF/00Z										
Pressure (hPa)		**U-comp (m/s)**				**V-comp (m/s)**				
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**	
150	225−100	590	−0.15	3.70	4.85	590	0.14	3.50	4.51	
300	425−225	458	−0.88	3.21	4.25	458	0.85	3.73	4.92	
500	625−425	333	−0.55	3.05	3.97	333	0.51	2.85	3.80	
700	775−625	272	−0.13	2.17	2.99	272	−0.52	2.32	3.08	
850	875−775	203	−0.30	2.32	3.13	203	−0.21	2.56	3.40	
900	910−875	1	1.98	1.98	1.98	1	−3.24	3.24	3.24	
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA	

WRF/00Z											
						Wind Direction (deg)					
Pressure (hPa)		**Wind Speed (m/s)**				**ROW_MEAN**			**AGGR**		
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**TOTAL**	**ME**	**MAE**
150	225−100	590	−0.52	4.02	5.17	30	0.18	3.49	590	0.47	NA
300	425−225	458	−0.98	3.52	4.63	30	2.36	5.25	457	2.65	NA
500	625−425	333	−0.22	2.92	3.78	30	2.74	6.00	333	2.86	NA
700	775−625	272	−0.65	2.28	3.02	30	−2.90	11.00	269	−3.64	NA
850	875−775	203	−0.09	2.41	3.12	30	−13.68	26.20	197	−4.40	NA
900	910−875	1	1.62	1.62	1.62	1	63.67	63.67	1	63.67	NA
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table A-11. Upper-air error statistics for meteorological variables for 12-km NAM forecast valid 0000 UTC, Domain 1.

NAM/00Z										
Pressure (hPa)		**Temperature (K)**				**Dew Point (K)**				
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**	
150	225−100	1074	0.43	1.66	2.13	NA	NA	NA	NA	
300	425−225	582	1.19	1.40	1.87	NA	NA	NA	NA	
500	625−425	588	0.31	0.83	1.12	NA	NA	NA	NA	
700	775−625	270	−0.12	0.81	1.09	NA	NA	NA	NA	
850	875−775	202	−1.62	1.99	2.48	NA	NA	NA	NA	
900	910−875	1	−3.85	3.85	3.85	NA	NA	NA	NA	
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA	
NAM/00Z										
Pressure (hPa)		**Rel Humidity (%)**				**Height (m)**				
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**	
150	225−100	NA	NA	NA	NA	514	21.52	21.99	25.53	
300	425−225	274	20.56	23.52	28.17	419	6.24	11.34	15.73	
500	625−425	575	10.47	19.12	24.46	333	−0.12	6.21	8.12	
700	775−625	270	8.64	13.67	16.77	273	−2.80	4.84	6.20	
850	875−775	202	10.70	11.92	13.82	234	−3.99	6.29	7.86	
900	910−875	1	9.60	9.60	9.60	1	−8.21	8.21	8.21	
1000	1010−910	NA	NA	NA	NA	118	−12.19	13.38	16.75	
NAM/00Z										
Pressure (hPa)		**U-comp (m/s)**				**V-comp (m/s)**				
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**RMSE**	
150	225−100	590	−0.32	3.63	4.72	590	0.23	3.69	4.69	
300	425−225	458	−0.70	3.16	4.23	458	0.68	3.83	5.13	
500	625−425	333	−0.62	3.23	4.40	333	0.43	2.81	3.71	
700	775−625	272	−0.36	2.19	3.07	272	−0.69	2.45	3.21	
850	875−775	203	−0.36	2.23	2.87	203	0.17	2.46	3.15	
900	910−875	1	1.78	1.78	1.78	1	−0.24	0.24	0.24	
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA	
NAM/00Z										
						Wind Direction (deg)				
Pressure (hPa)		**Wind Speed (m/s)**				**ROW_MEAN**				**AGGR**
CATEGORY	**RANGE**	**TOTAL**	**ME**	**MAE**	**RMSE**	**TOTAL**	**ME**	**MAE**	**TOTAL**	**ME**	**MAE**
150	225−100	NA	NA	NA	NA	30	0.63	3.87	590	0.87	NA
300	425−225	NA	NA	NA	NA	30	1.86	4.97	457	2.11	NA
500	625−425	NA	NA	NA	NA	30	2.74	5.72	333	2.73	NA
700	775−625	NA	NA	NA	NA	30	−3.25	10.56	269	−3.80	NA
850	875−775	NA	NA	NA	NA	30	−12.16	22.35	197	3.28	NA
900	910−875	NA	NA	NA	NA	1	40.82	40.82	1	40.82	NA
1000	1010−910	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

List of Symbols, Abbreviations, and Acronyms

AFWA	Air Force Weather Agency
ARL	U.S. Army Research Laboratory
ARW	Advanced Research WRF
BED	Battlefield Environment Division
CONUS	Continental United States
DoD	Department of Defense
DPG	Dugway Proving Ground
DTC	Developmental Testbed Center
FDDA	Four Dimensional Data Assimilation
GRIB	Gridded Binary Format
IDV	Integrated Data Viewer
MADIS	Meteorological Assimilation Data Ingest System
MAE	Mean Absolute Error
ME	Mean Error
MET	Model Evaluation Tools
METAR	Meteorological Terminal Aviation Weather Report
MODE	Method for Object-based Diagnostic Evaluation
NAM	North American Mesoscale Model
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NOAA	National Oceanic and Atmospheric Administration
NWP	Numerical Weather Prediction
RMSE	Root Mean Square Error
RRTM	Rapid Radiative Transfer Model
UTC	Coordinated Universal Time
WPPV3	WRF Post Processor Version 3
WRF	Weather Research and Forecasting
YSU	Yonsei State University

No. of
Copies **Organization**

1 (PDF) ADMNSTR
DEFNS TECHL INFO CTR
DTIC OCP
8725 JOHN J KINGMAN RD STE
0944
FT BELVOIR VA 22060-6218

3 HCs US ARMY RSRCH LAB
ATTN RDRL CIO LT
TECHL PUB
ATTN RDRL CIO LL
TECHL LIB
ATTN IMNE ALC HRR
MAIL & RECORDS MGMT
2800 POWDER MILL ROAD
ADELPHI MD 20783-1197

17 CDs US ARMY RSRCH LAB
ATTN RDRL CIE M
J PASSNER (3 COPIES)
Y RABY (3 COPIES)
J RABY (3 COPIES)
G VAUCHER (3 COPIES)
D KNAPP
R FLANIGAN
S KIRBY
B DUMAIS
T JAMESON

1 CD US ARMY RSRCH LAB
ATTN RDRL CIE
P CLARK

1 CD US ARMY RSRCH LAB
ATTN RDRL CIE D
D HOOCK

3 CDs DR J MCLAY
NAVAL RESEARCH LABORATORY
7 GRACE HOPPER AVE STOP 2
MONTEREY CA 93943

No. of
Copies **Organization**

1 CD R CRAIG DAF CIVILIAN
HQ AFWA 2WXG 16WS/WXN
101 NELSON DRIVE
OFFUTT AFB, NE 68113-1023

3 CDs ARMY JOINT SUPPORT TEAM
SFAE IEW&S DCGS A
ATTN G BARNES
238 HARSTON ST BLDG 90060
HURLBURT FIELD FL 32544

1 CD T FOWLER
NCAR DEVELOPMENTAL
TESTBED CENTER
P.O. BOX 3000
BOULDER CO 80307-3000

1 CD J H GOTWAY
NCAR DEVELOPMENTAL
TESTBED CENTER
PO BOX 3000
BOULDER CO 80307-3000

1 CD J STALEY
ARMY WEATHER PROPONENT
OFFICE
INTEGRATION
SYNCHRONIZATION AND
ANALYSIS (CDID)
US ARMY INTELLIGENCE CENTER
OF EXCELLENCE
550 CIBEQUE ST BLDG 61730
FT HUACHUCA AZ 85613

Total: 33 (1 PDF, 29 CDs, 3 HCs)