WRFv3.1.1+ QNSE Test and Evaluation

Jamie Wolff*, Louisa Nance, John Halley Gotway, Paul Oldenberg

National Center for Atmospheric Research, Boulder, CO

1. Introduction

The Weather Research and Forecasting (WRF) model is a state-of-the-art numerical weather prediction system that is highly configurable and suitable for a broad range of weather applications. Given the numerous options available, it is important to rigorously test configurations to assess the performance of select configurations for specific applications. The Air Force Weather Agency (AFWA) is interested in improvements in the characterization of the planetary boundary layer (PBL) and surface layer. The Quasi-Normal Scale Elimination (QNSE) PBL and surface layer schemes developed by Sukoriansky, Galperin and Perov, (Sukoriansky et al. 2005) are new features available since WRF version 3.1 with the goal of addressing these issues. To assess the performance of these new schemes, the Developmental Testbed Center (DTC) performed testing and evaluation with the Advanced Research WRF (ARW) dynamic core (Skamarock et al. 2008) for two physics suite configurations at the request of the sponsor, AFWA. One configuration was based on AFWA's Operational Configuration, which now provides a baseline for testing and evaluating new options available in the WRF system. The second configuration substituted AFWA's current operational PBL and surface layer schemes with the QNSE schemes. Forecast verification statistics were computed for the two configurations and the analysis was based on the objective statistics of the model output.

2. Experiment Design

The end-to-end forecast system employed the WRF Preprocessing System (WPS), WRF, WRF Postprocessor (WPP) and graphics generation using NCL. Post-processed forecasts were verified using the Model Evaluation Tools (MET). In addition, the full data set was archived and made available for dissemination. The codes utilized were based on the official released versions of WPS (v3.1.1), WPP (v3.1), and MET (v2.0). Both WPP and MET included relevant bug fixes that were checked into the respective code repositories prior to testing. For WRF, a tag from the repository was also used, which was based on v3.1.1 with a considerable number of updates.

*Corresponding author email address: jwolff@ucar.edu

2.1 Forecast Periods

Forecasts were initialized every 36 hours from 2 June 2008 through 31 May 2009, automatically creating a combination of initialization times including both 00 and 12 UTC, for a total of 243 cases (see Appendix A for a list of the cases). The forecasts were run out to 48 hours with output files generated every 3 hours.

2.2 Initial and Boundary Conditions

Initial conditions (ICs) and lateral boundary conditions (LBCs) were derived from the 0.5° x 0.5° Global Forecast System (GFS). Output from AFWA's Agricultural Meteorological Modeling (AGRMET) System was utilized for the lower boundary conditions (LoBCs) in addition to a daily, real-time sea surface temperature product from Fleet Numerical Meteorology and Oceanography Center (FNMOC), which was used to initialize the sea surface temperature (SST) field for the forecasts. Finally, the time-invariant components of the LoBCs (topography, soil and vegetation type etc.) were derived from United States Geological Survey (USGS) input data.

2.3 Model Configuration Specifics

2.3.1 Domain Configuration

A 15-km contiguous U.S. (CONUS) grid was employed in this test. The domain (Fig. 1) was selected such that it covers complex terrain, plains, and coastal regions spanning from the Gulf of Mexico, north, to Central Canada in order to capture diverse regional effects for worldwide comparability. The domain was 403 x 302 gridpoints, for a total of 121,706 gridpoints. The Lambert-Conformal map projection was used and the model was configured to have 56 vertical levels (57 sigma entries) with the model top at 10 hPa.

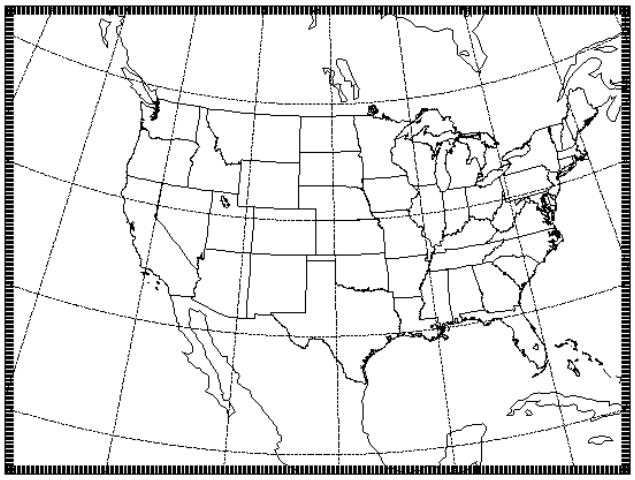

Figure 1. Map showing the boundary of the WRF-ARW computational domain.

2.3.2 Other Aspects of Model Configuration

The two physics suite configurations used for each model configuration in this test are described in the table below. The model configuration based on AFWA's Operational Configuration will be referred to as AFWA, while the companion configuration will be referred to as QNSE.

	Current AFWA Config (AFWA)	QNSE replacement (QNSE)
Microphysics	WRF Single-Moment 5 scheme	WRF Single-Moment 5 scheme
Radiation SW and LW	Dudhia/RRTM schemes	Dudhia/RRTM schemes
Surface Layer	Monin-Obukhov similarity theory	QNSE
Land-Surface Model	Noah	Noah
Planetary Boundary Layer	Yonsei University scheme	QNSE
Convection	Kain-Fritsch scheme	Kain-Fritsch scheme

Both configurations were run with a long timestep of 90 s, and an acoustic step of 4 was used. Calls to the boundary layer and microphysics were performed every time step, whereas the cumulus parameterization was called every 5 minutes. Radiation was called every 30 minutes.

3. Model Verification

Objective model verification statistics were generated using the MET package. MET is comprised of grid-to-point comparisons, which were utilized to compare gridded surface and upper-air model data to point observations, as well as grid-to-grid comparisons, which were utilized to verify QPF. Verification statistics generated by MET for each retrospective case were used to compute and plot specified aggregated statistics using routines developed by the DTC in the statistical programming language, R.

Though several domains were verified for the surface and upper air, as well as precipitation variables, only the CONUS domain is described in detail for this paper. In addition to the regional area stratification, the verification statistics were also stratified by vertical level, forecast lead time and/or precipitation threshold. The annual aggregations only will be described for this paper. A complete set of results for all sub-domains and seasonal aggregations are available on the DTC website (http://verif.rap.ucar.edu/eval/afwa_rc/).

Each type of verification metric is accompanied by confidence intervals (CIs) at the 99% level, computed using the appropriate statistical method. Both configurations were run for the same cases allowing for a pair-wise difference methodology to be applied, as appropriate. The CIs on the pair-wise differences between statistics for the two configurations objectively determines whether the differences are statistically significant (SS); if the CIs on the pair-wise verification statistics include zero the differences are not statistically significant. Because frequency bias is not amenable to a pair-wise difference comparison due to the nonlinear attributes of this metric, the more powerful method to establish SS could not be used and, thus, a more conservative estimate of SS was employed based solely on whether the aggregate statistics, with the accompanying CIs, overlapped between the two configurations. If no overlap was noted for a particular threshold, the differences between the two configurations were considered SS.

3.1 Temperature, Dew Point Temperature, and Winds

Objective model verification statistics were generated for surface (using METAR and buoy observations) and upper air (using RAOBS) temperature, dew point temperature, and wind. Because shelter-level variables are not realistic at the initial model time, surface verification results start at the 3-hour lead time and go out 48 hours by 3-hour increments. For upper air, verification statistics were computed at the mandatory levels using radiosonde observations and computed at 12-hour intervals out to 48 hours. Because of known errors associated with radiosonde moisture measurements at high altitudes, the analysis of the upper air dew point temperature verification focuses on levels at and below 500 hPa. Bias and bias-corrected root-mean-square-error (BCRMSE) were computed separately for surface and upper air observations. The CIs were computed from the standard error estimates about the median value of the stratified results for the surface and upper air statistics of temperature, dew point temperature and wind using a parametric method and a correction for first-order autocorrelation.

3.2 Precipitation

For the QPF verification, a grid-to-grid comparison was made by first interpolating the precipitation analyses to the 15-km model integration domain. Accumulation periods of 3 and 24 hours were examined. The observational datasets used were the NCEP Stage II analysis for the 3-hour accumulation and the NCEP/CPC daily gauge analysis for the 24-hour accumulation. Because the 24-hour accumulation observations are only valid at 12 UTC, the 24-hour QPF were examined for the 24- and 48-hour lead times for the 12 UTC initializations and 36-hour lead time for the 00 UTC initializations. Traditional verification metrics computed included the frequency bias and the equitable threat score, or Gilbert skill score (GSS). For the precipitation statistics, a bootstrapping CI method was applied.

4. Verification Results

Differences are computed between the two configurations by subtracting the QNSE configuration from the AFWA configuration. BCRMSE is always a positive quantity and a perfect score is zero. Given these properties, differences that are negative (positive) indicate the AFWA (QNSE) configuration has a lower BCRMSE. For GSS, the perfect score is one and the no-skill forecast is zero. Thus, if the pair-wise difference is positive (negative) the AFWA (QNSE) configuration has a higher GSS. The properties of bias (which has a perfect score of zero) and frequency bias (which has a perfect score of one) are not as conducive to generalized statements such as those that can be made for BCRMSE and GSS. Both of these metrics can have positive or negative values. Given this, when looking at the pair-wise differences it is important to also note the magnitude of the bias in relation to the perfect score for each individual configuration to know which configuration has a smaller bias.

4.1 Upper Air

4.1.1 Temperature BCRMSE and bias

The overall distribution for temperature BCRMSE for both the AFWA and QNSE configurations show a minimum error between 500 and 300 hPa and, as expected, the BCRMSE increases with forecast lead time (48-hr lead time shown in Fig. 2). The pair-wise differences for the annual aggregation at all forecast lead times indicate all SS differences at and below 400 hPa, as well as those at and above 150 hPa, favor the AFWA configuration. Conversely, the SS pair-wise differences at 200 and 300 hPa favor the QNSE configuration. It is worth noting, however, that the relative magnitudes of the SS pair-wise differences favoring the AFWA configuration are larger than those favoring the QNSE configuration.

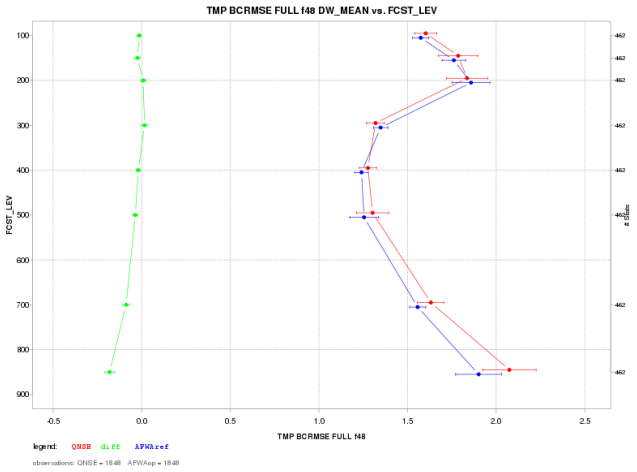

Figure 2. Vertical profile of the median BCRMSE for temperature (C) for the full integration domain aggregated across the entire year of cases (annual) for the 48-hour lead time. The AFWA configuration is shown in blue, the QNSE configuration in red, and the differences (AFWA-QNSE) in green. The horizontal bars represent the 99% CIs.

Both configurations produce a temperature bias that transitions from cold at lower levels to warm at upper levels. The level at which this transition occurs varies slightly with lead time (only 48-hour lead time shown in Fig. 3). The SS pair-wise differences indicate the QNSE configuration tends to produce the smallest temperature bias at 850 hPa and 200 hPa, whereas the AFWA configuration tends to produce the smallest bias between 700 and 300 hPa.

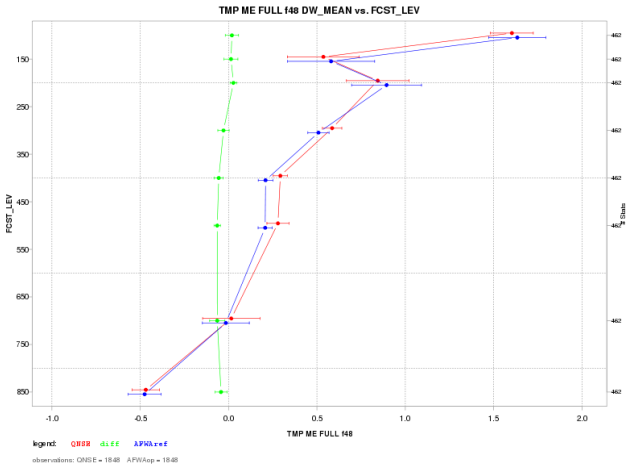

Figure 3. Vertical profile of the median bias for temperature (C) for the full integration domain aggregated across the entire year of cases for the 48-hour lead time. The AFWA configuration is shown in blue, the QNSE configuration in red, and the differences (AFWA-QNSE) in green. The horizontal bars represent the 99% CIs.

4.1.2 Dew Point Temperature BCRMSE and bias

The dew point temperature BCRMSE increases as the pressure decreases for both configurations and gradually increases with increasing lead time (Fig. 4). All SS pair-wise differences for the pair-wise comparison correspond to the AFWA configuration having a lower BCRMSE.

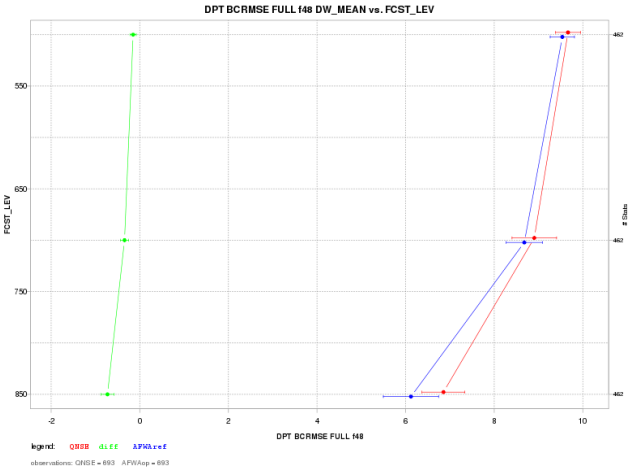

Figure 4. Vertical profile of the median BCRMSE for dew point temperature (C) for the full integration domain aggregated across the entire year of cases (annual) for the 48-hour lead time. The AFWA configuration is shown in blue, the QNSE configuration in red, and the differences (AFWA-QNSE) in green. The horizontal bars represent the 99% CIs.

Both configurations tend to produce a positive dew point temperature or moist bias at all levels and lead times for the annual aggregation (not shown). The magnitude of the bias is fairly consistent and actually decreases slightly for the longer lead times. The SS pair-wise differences generally favor the QNSE configuration.

4.1.3 Wind BCRMSE and bias

The vertical distribution of vector wind BCRMSE for both configurations exhibits the same general properties for all lead times. The distribution increases to a maximum between 300 and 200 hPa and then decreases aloft (Fig. 5). All SS pair-wise differences correspond to the AFWA configuration having smaller errors then the QNSE configuration regardless of level, or lead time.

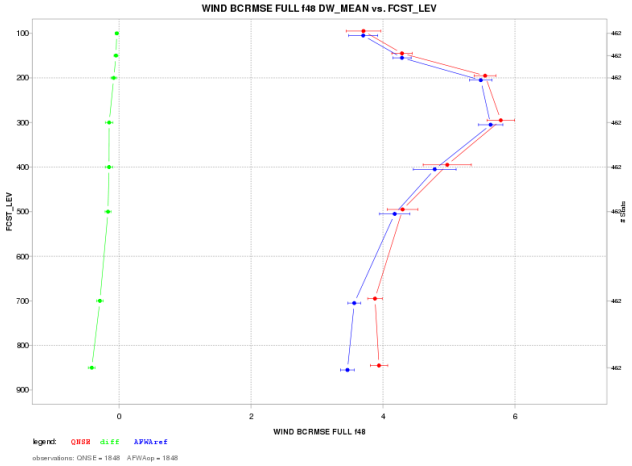

Figure 5. Vertical profile of the median BCRMSE of vector winds (m/s) for the full integration domain at the 48-hour lead time aggregated across the entire year of cases. The AFWA configuration is shown in blue, the QNSE configuration in red, and the differences (AFWA-QNSE) in green. The horizontal bars represent the 99% CIs.

Vertical profiles of wind speed bias indicate the winds for the AFWA configuration are non-biased at 850 hPa, whereas the winds for the QNSE configuration are too strong (Fig. 6). The wind speed bias for both configurations transitions to winds that are too light at upper levels. For this metric, the QNSE configuration has a consistent SS bias towards higher wind speeds as compared to the AFWA configuration at all levels below 400 hPa. This translates to the QNSE configuration having SS smaller bias when the overall wind speed bias is too light (at and above 700 hPa) and the AFWA configuration having SS smaller bias at levels where the overall wind speed bias is too fast (generally, below 850 hPa).

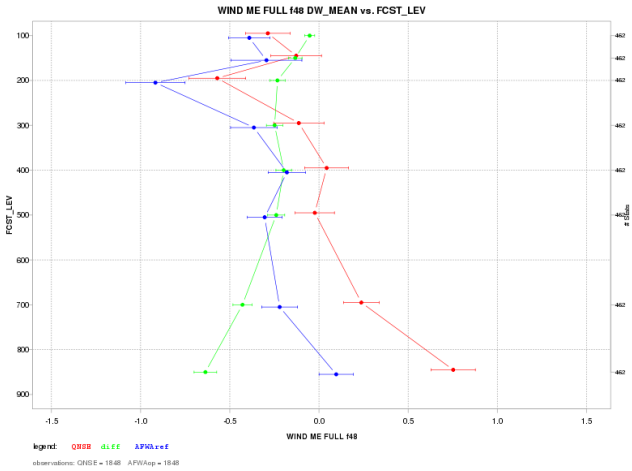

Figure 6. Vertical profile of the median bias of wind speed (m/s) for the full integration domain aggregated across the entire year of cases (annual)

for the 48-hour lead time. The AFWA configuration is shown in blue, the QNSE configuration in red, and the differences (AFWA-QNSE) in green. The horizontal bars represent the 99% CIs.

4.2 Surface

Following the completion of the extensive testing on the comprehensive set of cases, the developers of the QNSE scheme uncovered a bug in the code (based on preliminary results from one month of testing provided by the DTC) leading to a significant misrepresentation of surface (2m and 10m diagnostic) fields only, the upper air results remain unaffected. Because of the late date of this discovery, it was not feasible for the DTC to rerun after a fix had been checked into the WRF repository. All 2m and 10m diagnostic fields contain this known bug and, in the interest of space, will not be discussed for this paper (but can be found on the DTC website). A rerun of this configuration will be conducted with WRFv3.2 to assess the exact impact of the bug and evaluate the latest QNSE scheme.

4.2.4 3-hourly QPF GSS and bias

When evaluating the GSS for precipitation it is important to know the number of observations that make up a particular distribution of values for each threshold. The base rate, indicating the ratio of observed grid box events to the total number of grid boxes in the domain, is shown on each precipitation plot by threshold. As the base rate decreases, the number of cases observed decreases and the event becomes infrequent. With this decreasing base rate is often an increase in the size of the CIs as well, indicating more spread and less confidence in the median value.

When examining the GSS values for the 3-hour QPF, it is seen that the highest GSS values occur at the lowest precipitation threshold of 0.01" and steadily decrease to near-zero for thresholds greater than 1.0" (Fig. 7). The number of observed events by threshold has a similar trend. The base rate for the 00 UTC 12-hour forecast is lower than the 12 UTC 12-hour forecast, likely due to the increased precipitation potential in the late afternoon with the heating cycle. In the analysis presented here, no SS pair-wise differences are noted.

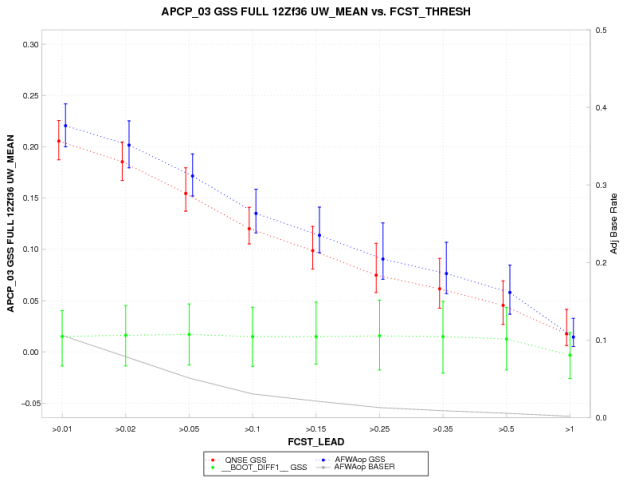

Figure 7. Threshold series plot of 3-hour accumulated precipitation (in) for median GSS for the 12 UTC initializations aggregated across the entire year of cases for the 36-hour lead time. The AFWA configuration is shown in blue, the QNSE configuration in red, and the differences (AFWA-QNSE) in green. The vertical bars represent the 99% CIs. Associated with the second y-axis, the light grey line is the adjusted base rate, or the ratio of observed grid box events to the total number of grid boxes in the domain, by threshold.

With few exceptions, both configurations have a SS high bias for thresholds less than 0.25" regardless of initialization time (Fig. 8). Above 0.25" the general trend is a decreasing bias where in many cases the CIs encompass one (perfect score for frequency bias) for the 0.35" threshold and then transition to a SS low bias for higher thresholds. SS differences are generally noted for the lowest thresholds from forecasts valid at 00 UTC, regardless of the initialization time.

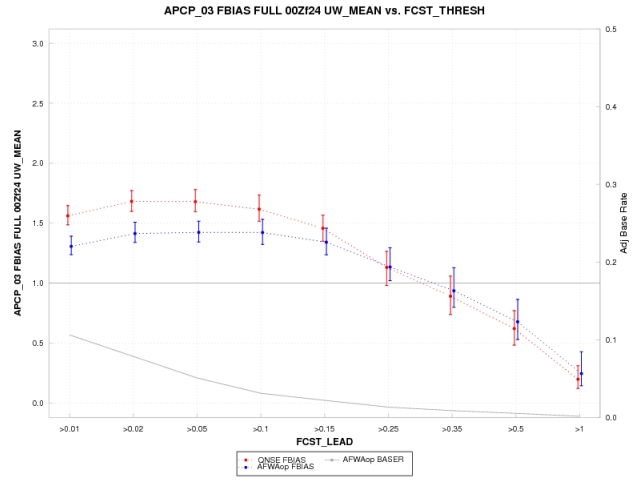

Figure 8. Threshold series plot of 3-hour precipitation accumulation (in) for median frequency bias for the 00 UTC initializations 24-hour lead time only aggregated across the entire year of cases. The AFWA configuration is shown in blue and the QNSE configuration in red. The vertical bars represent the 99% CIs. Associated with the second y-axis, the light grey line is the adjusted base rate, or the ratio of observed grid box events to the total number of grid boxes in the domain, by threshold.

4.2.5 Daily Precipitation GSS and bias

The base rate for the 24-hour QPF is over 30% for the lowest threshold but the decrease in GSS values as the threshold increases is similar to that shown for the 3-hour QPF (not shown). No SS pair-wise differences are seen for any lead time or threshold.

The overall magnitude of the 24-hour accumulation biases for the 00 and 12 UTC initializations are similar up to the 1" threshold, and reveal a general SS high bias for both configurations (Fig. 9). For the largest accumulation thresholds (greater than 1.5" or 2") the CIs are very large (encompassing one) and are, therefore, classified as nonbiased due to low confidence in the actual magnitude or sign of the bias. Once again, when using the more conservative method for assessing SS between the two configurations all favor the AFWA configuration and occur at the lowest thresholds.

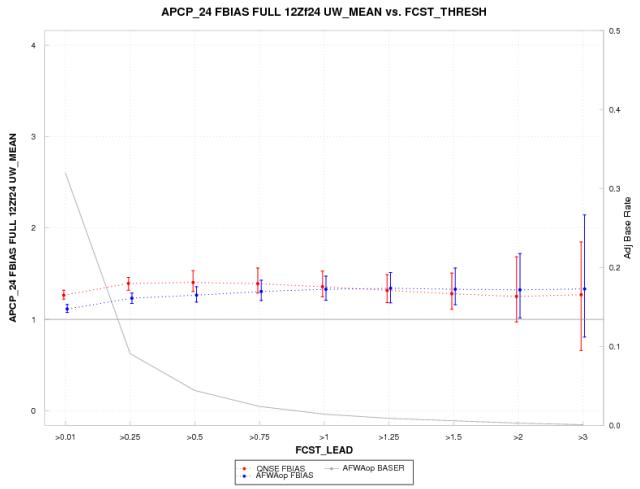

Figure 9. Threshold series plot of 24-hour precipitation accumulation (in) for median frequency bias for the 12 UTC initializations 24-hour lead time only aggregated across the entire year of cases. The AFWA configuration is shown in blue and the QNSE configuration in red. The vertical bars represent the 99% CIs. Associated with the second y-axis, the light grey line is the adjusted base rate, or the ratio of observed grid box events to the total number of grid boxes in the domain, by threshold.

5. Summary

Two WRF-ARW configurations were comprehensively tested and evaluated to assess the impact of the new QNSE PBL and surface layer schemes available in WRF, using AFWA's Operational Configuration as a baseline. Because both configurations were run for the same cases, pair-wise differences were computed for standard verification metrics between the two configurations, and an assessment of the statistical significance (SS) was included. In general, the AFWA configuration was favored more often than the QNSE configuration. However, for some metrics and certain levels, lead times, or thresholds, QNSE was favored. It may be noted, though, that the relative magnitudes of the SS differences favoring the AFWA configuration are generally larger than those favoring the QNSE configuration.

Please see: http://verif.rap.ucar.edu/eval/afwa_rc/ for full details and results of this test and evaluation project.

6. References

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang and J. G. Powers, 2008: A Description of the Advanced Research WRF Version 3, NCAR Tech Note, NCAR/TN–475+STR, 113 pp.

Sukoriansky, S., B. Galperin, and V. Perov, 2005: Application of a new spectral theory of stably stratified turbulence to the atmospheric boundary layer over sea ice. *Boundary-Layer Meteorol.,* **117,** 231-257.

Acknowledgements: The DTC is funded by the National Oceanic and Atmospheric Administration, the Air Force Weather Agency, and National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National Science Foundation.