P470 **WRF QNSE TEST AND EVALUATION**

Jamie Wolff*, Louisa Nance, John Halley Gotway, Paul Oldenburg, Michelle Harrold, and Zach Trabold

National Center for Atmospheric Research/Research Applications Laboratory (NCAR/RAL) and Developmental Testbed Center (DTC), Boulder, CO USA

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. Initial testing of both configurations was conducted using code based on WRF v3.1.1, and a retest of both configurations was conducted in exactly the same manner as the initial testing with the most recent WRF v3.2.1. Forecast verification statistics were computed for the two configurations and the two versions of the code, 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 for these tests were based on the official released versions of WPS (v3.1.1 and v3.2.1), WPP (v3.1 and v3.2), and MET (v2.0 and v3.0). 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 for the first test, which was based on v3.1.1 with a considerable number of updates, while the retest used officially release code (v3.2.1).

*Corresponding author address: Jamie Wolff, NCAR/RAL, P.O. Box 3000, Boulder, CO 80307, email: jwolff@ucar.edu

2.1 Forecast Periods

Forecasts were initialized every 36 hours from 2 June 2008 through 31 May 2009, by design creating a combination of initialization times including both 00 and 12 UTC, for a total of 243 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 for these tests. 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.

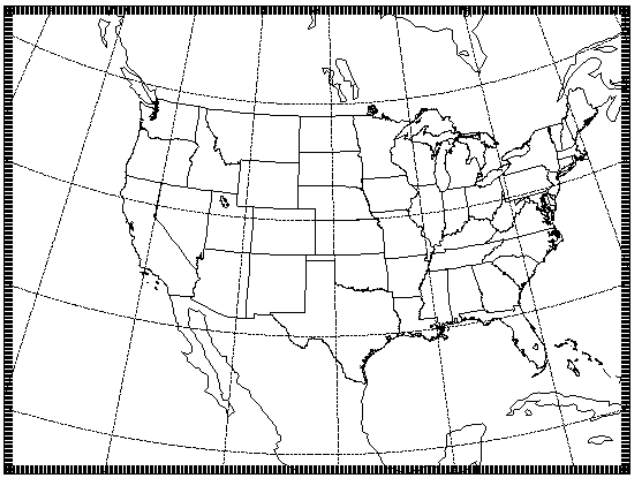

Figure 1. Map showing the boundary of the WRF-ARW computational domain.

2.3.2 Other Aspects of Model Configuration

The two physics suite configurations used for each model configuration in these tests 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 model quantitative precipitation forecast (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 variables and precipitation, 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 here. 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_test).

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 and two model versions 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.

Because of the large dataset available, allowing for a good approximation of the distribution of results, a large number SS pair-wise differences was anticipated. However, in many cases the SS pair-wise differences were not practically meaningful. Thus, to establish practical significance (PS), the data was censored to only look at SS pair-wise differences that were greater than the operational measurement uncertainty requirements and instrument performance as defined by the World Meteorological Organization (WMO) (http://www.wmo.int/pages/prog/www/IMOP/publications/CIMO-Guide/1st-Suppl-to-7th_draft/pdf/Annex_I_1B.pdf). The following criteria were applied to determine PS pair-wise differences between the configurations and versions for each variable: i) temperature and dew point temperature differences greater than 0.1K, ii) wind speed differences greater than 0.5 ms^{-1}, iii) and precipitation differences greater than 0.1 mm.

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 were 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 National Centers for Environmental Prediction (NCEP) Stage II analysis for the 3-hour accumulation and the NCEP/Climate Prediction Center (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 for the same version of the code by subtracting the QNSE configuration from the AFWA configuration or between two versions of the code for one configuration by subtracting v3.1.1 from v3.2.1. BCRMSE is always a positive quantity, and a perfect score is zero. Given these properties, differences that are negative (positive) indicate the AFWA or v3.2.1 (QNSE or v3.1.1)

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 or v3.2.1 (QNSE or v3.1.1) 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.

Overall, the distributions between the configuration differences for v3.1.1 and v3.2.1 are similar so only v3.2.1 will be described in detail here, with major differences compared to v3.1.1 highlighted as needed. In addition, the changes in the objective verification results due only to an updated version of WRF for the AFWA configuration will be discussed.

4.1 Upper Air

4.1.1 Temperature BCRMSE and Bias

The overall distribution for temperature BCRMSE for both the AFWA and QNSE configurations for v3.2.1 show a minimum error between 500 and 300 hPa and maxima at 850 and 200 hPa. The shape of the distribution remains the same; however, the BCRMSE values increase 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 (Appendix A, Table 1). Conversely, the SS pair-wise differences at 200 and 300 hPa favor the QNSE configuration. However, PS differences are only noted at 850 hPa for lead times over 12 hours and at 700 hPa for the 48 hour lead time, all of which favor the AFWA configuration. For the AFWA version difference, no PS differences are seen for upper air temperature BCRMSE (Appendix B, Table 2).

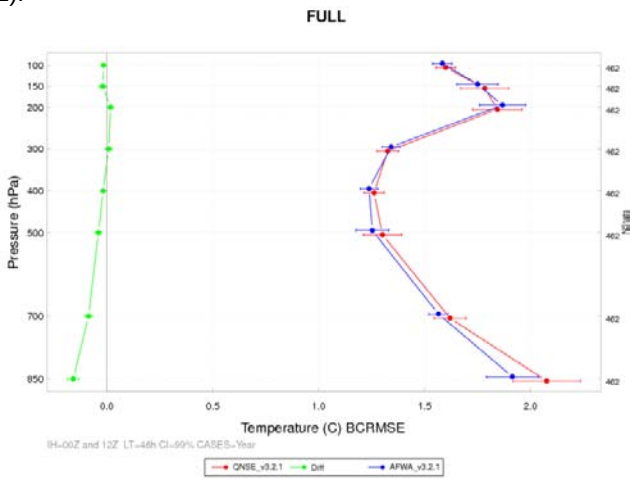

Figure 2. Vertical profile of the median BCRMSE for temperature (°C) for the full integration domain aggregated across the entire year of cases for the 48-hour lead time. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 configuration in red, and the differences (AFWA-QNSE) in green. The 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 (Fig. 3). Several SS pair-wise differences are noted; however, none are PS between the AFWA and QNSE configurations for v3.2.1. When examining the version differences for the AFWA configuration, differences at the 200 and 150 hPa levels are PS and favor v3.2.1.

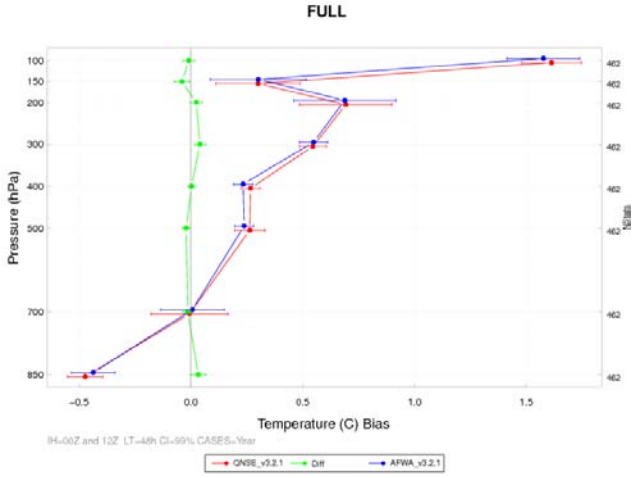

Figure 3. Vertical profile of the median bias for temperature (C) for the full integration domain aggregated across the entire year of cases for the 48-hour lead time. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 configuration in red, and the differences (AFWA-QNSE) in green. The horizontal 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 with v3.2.1 and gradually increases with increasing lead time (Fig. 4). At all lead times and levels, the AFWA configuration has lower BCRMSE values that are not only SS, but also PS. No PS differences are noted for the AFWA version difference comparison.

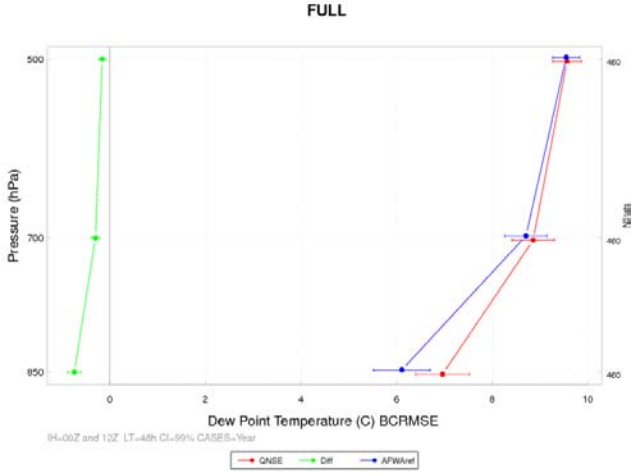

Figure 4. Vertical profile of the median BCRMSE for dew point temperature (C) for the full integration domain aggregated across the entire year of cases for the 48-hour lead time. The

AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 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. Converse to BCRMSE, the pair-wise differences generally favor the QNSE configuration and are PS in most cases.

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 gradually to a maximum between 300 and 200 hPa and then decreases aloft (Fig. 5). None of the SS pair-wise differences favoring the AFWA configuration are PS.

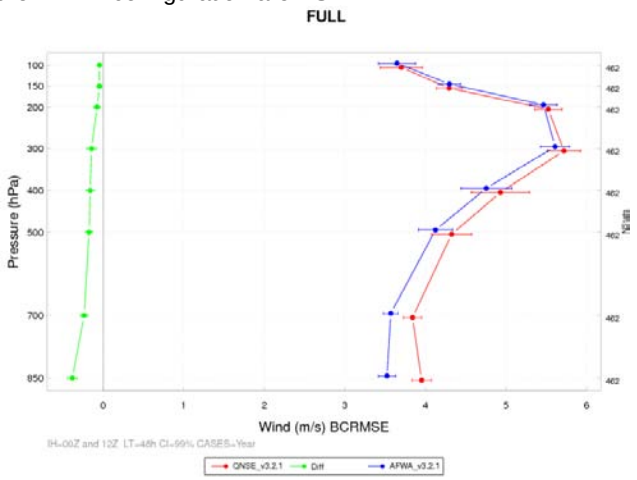

Figure 5. Vertical profile of the median BCRMSE of vector winds (m/s) for the full integration domain at the 48-hour lead time aggregated across the entire year of cases. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 configuration in red, and the differences (AFWA-QNSE) in green. The horizontal bars represent the 99% CIs.

Vertical profiles of wind speed bias indicate the winds for the AFWA configuration are generally too light, while for the QNSE configuration the winds transition from a high bias near the surface to a low bias at upper levels (Fig. 6). 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 generally being favored with SS lower bias at all levels and lead times at and above 500 hPa (when the overall wind speed bias is too low for both configurations), while the AFWA configuration is favored at 850 hPa (where the wind speed bias is too high for both configurations). The only PS differences, however, are those favoring the AFWA configuration at 850 hPa at lead times at and beyond 24 hours. As was the case for BCRMSE, none of the SS pair-wise differences favoring the AFWA configuration are PS for bias.

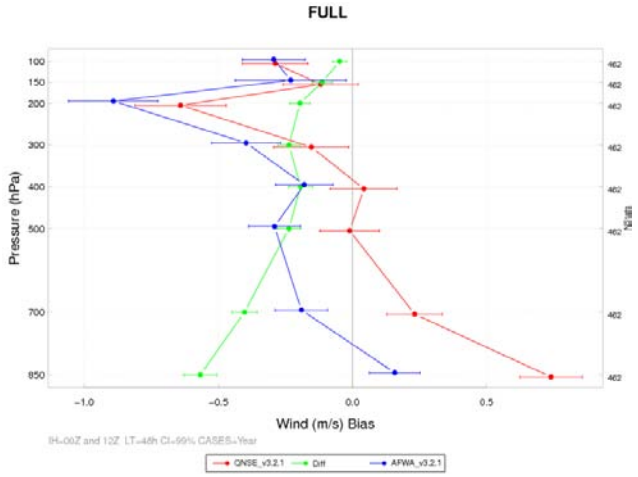

Figure 6. Vertical profile of the median bias of wind speed (m/s) for the full integration domain aggregated across the entire year of cases for the 48-hour lead time. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 configuration in red, and the differences (AFWA-QNSE) in green. The horizontal bars represent the 99% CIs.

4.2 Surface

4.2.1 Temperature BCRMSE and Bias

The surface temperature BCRMSE for both configurations undergoes a slight increase with lead time (Fig. 7). Diurnal variations are also evident with the lowest error values noted around valid times of 06-09 UTC and the maximum errors valid at 00 UTC. In all cases, SS pair-wise differences favor the AFWA configuration and most are also PS (Appendix A, Table 2). For the AFWA version differences, all SS pair-wise differences seen favor v3.1.1; however, none are PS (Appendix B, Table 2).

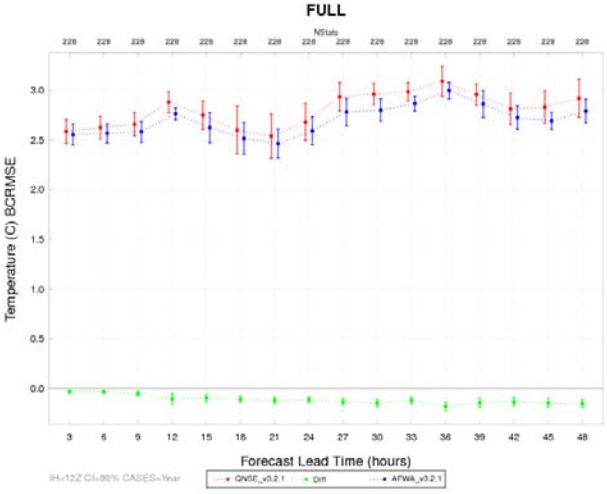

Figure 7. Time series plot of 2m AGL temperature (C) for median BCRMSE for the 12 UTC initializations only aggregated across the entire year of cases. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 configuration in red, and the differences (AFWA-QNSE) in green. The vertical bars represent the 99% CIs.

Time series plots of surface temperature bias exhibit a strong diurnal cycle for both configurations for v3.2.1. For the AFWA configuration, this cycle corresponds to a cold surface temperature bias during the daytime hours and a

warm bias during the overnight hours (Fig. 8), indicating the configuration is under-predicting the amplitude of the diurnal temperature cycle. The magnitude and the sign of the bias are dependent on the phase of the diurnal cycle but the amplitude does not increase with lead time. Conversely, the QNSE configuration produces a cold bias for all forecast lead times, where the magnitude of the bias is at 00 UTC and exhibits a slight overall increase with lead time. For the annual aggregation, the pair-wise differences are PS and favor the AFWA configuration at all lead times. When examining the AFWA version differences, the favored version is dependent on the seasonal aggregation, initialization time and lead time. In general, v3.2.1 is favored during the overnight hours, while v3.1.1 is favored during the daytime hours. While many SS pair-wise differences are seen, just over half of those are PS (most of which favor v3.2.1).

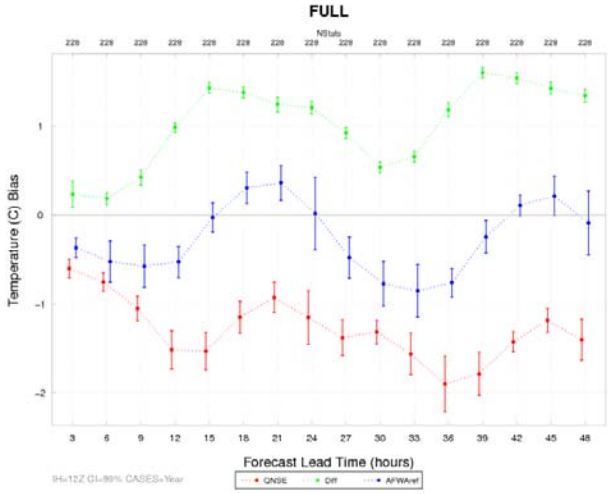

Figure 8. Time series plot of 2m AGL temperature (C) for median bias for the 12 UTC initializations only aggregated across the entire year of cases. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 configuration in red, and the differences (AFWA-QNSE) in green. The vertical bars represent the 99% CIs.

4.2.2 Dew Point Temperature BCRMSE and Bias

A general diurnal cycle similar to that of BCRMSE for temperature is noted for dew point temperature, with the maximum errors for both configurations with v3.2.1 occurring around 21-00 UTC and the minimum around 06 UTC (Fig. 9). PS pair-wise differences generally indicate that the QNSE configuration is favored for valid times around 06 UTC while the AFWA configuration is favored between 21-00 UTC. More PS pair-wise differences favoring the QNSE configuration are noted for the aggregation of the 00 UTC initializations (not shown) as compared to the 12 UTC initializations. There are several SS differences favoring v3.2.1 for the AFWA version difference comparison; however, none are PS.

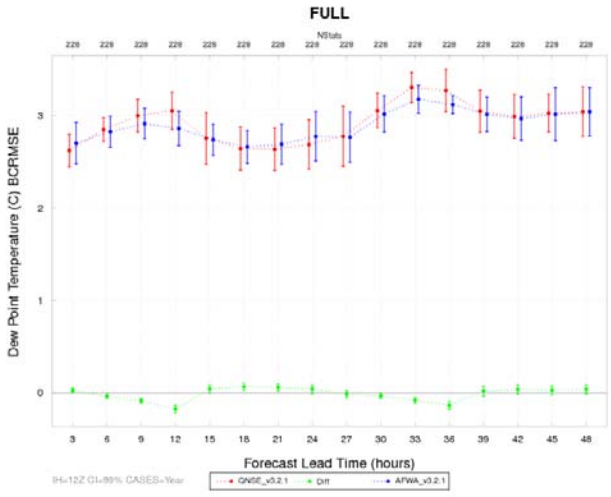

Figure 9. Time series plot of 2m AGL dew point temperature (C) for median BCRMSE for the 12 UTC initializations only aggregated across the entire year of cases. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 configuration in red, and the differences (AFWA-QNSE) in green. The vertical bars represent the 99% CIs.

The bias for surface dew point temperature from the QNSE configuration is consistently SS high for most lead times (Fig. 10). The AFWA configuration exhibits a SS high bias between 18-00 UTC and no distinguishable bias for all other lead times (CIs encompass zero). All pair-wise differences for this variable and metric are PS, with a majority favoring the AFWA configuration. When looking at the AFWA version differences, both versions perform better for certain lead times resulting in no clear, consistent favored version.

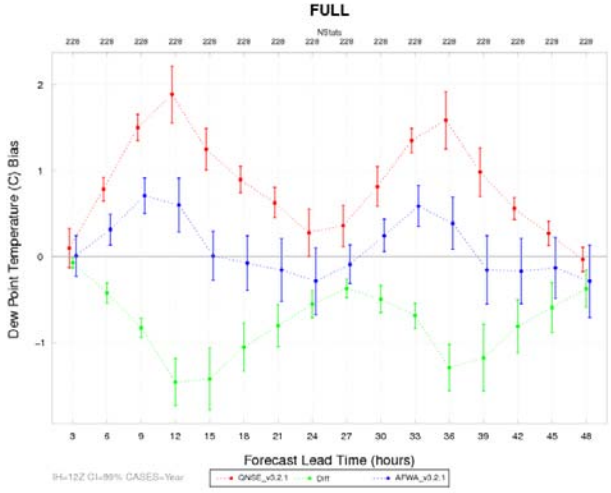

Figure 10. Time series plot of 2m AGL dew point temperature (C) for median bias for the 12 UTC initializations only aggregated across the entire year of cases. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 configuration in red, and the differences (AFWA-QNSE) in green. The vertical bars represent the 99% CIs.

4.2.3 Wind BCRMSE and Bias

The BCRMSE for the surface wind vectors show a weak diurnal signal and an overall slight increase in error with longer lead times for both configurations (Fig. 11). The largest wind vector errors occur around 00 UTC and the

smallest around 12 UTC. Again, several SS pair-wise differences are noted, most favoring the AFWA configuration; however, none are PS.

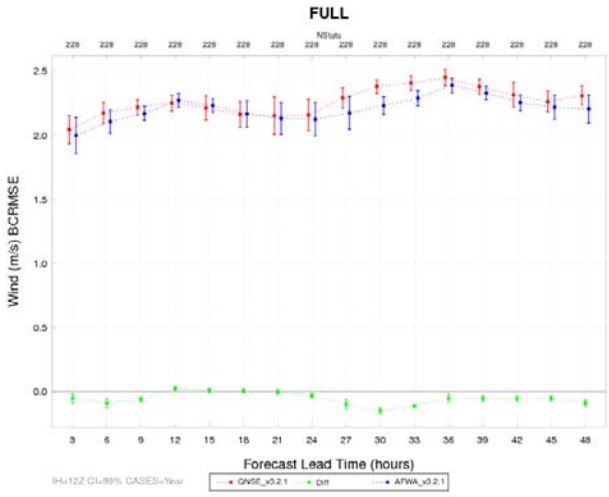

Figure 11. Time series plot of 2m AGL vector winds (m/s) for median BCRMSE for the 12 UTC initializations only aggregated across the entire year of cases. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 configuration in red, and the differences (AFWA-QNSE) in green. The vertical bars represent the 99% CIs.

Both configurations produce a high wind speed bias at the surface for all forecast lead times (Fig. 12). The AFWA configuration produces a larger magnitude diurnal cycle with biases that are larger than those for the QNSE configuration during the overnight hours and smaller during the daytime hours. This relationship leads to the general statement for the pair-wise differences that the QNSE has a SS smaller bias during the overnight hours (between about 00 and 12 UTC) and the AFWA configuration has a SS smaller bias during the daytime hours. However, it is noted that while nearly all differences favoring the AFWA configuration are PS, far fewer are PS when the QNSE configuration is favored.

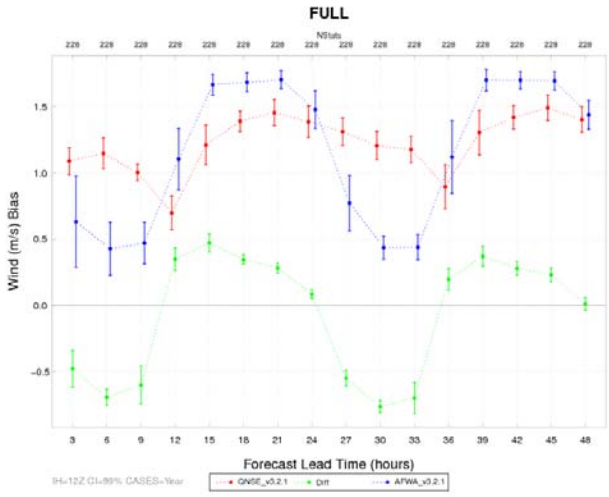

Figure 12. Time series plot of 2m AGL wind speed (m/s) for median bias for the 12 UTC initializations only aggregated across the entire year of cases. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 configuration in red, and the differences (AFWA-QNSE) in green. The vertical bars represent the 99% CIs.

4.2.4 3-hourly QPF GSS and Frequency 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. 13). 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, some SS pair-wise differences are noted; however, none are PS.

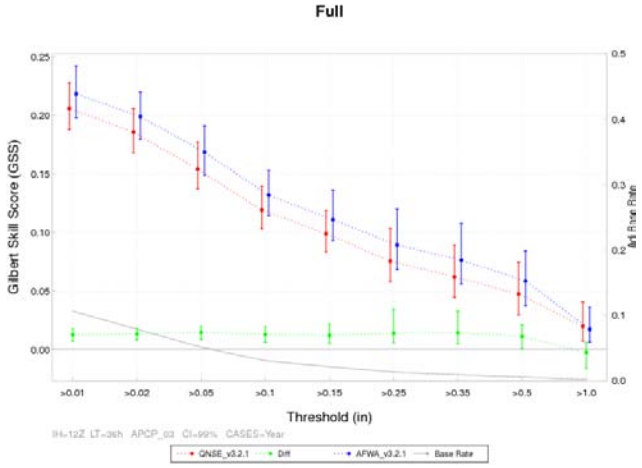

Figure 13. Threshold series plot of 3-hour accumulated precipitation (in) for median GSS for the 12 UTC initializations aggregated across the entire year of cases for the 36-hour lead time. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3,2,1 configuration in red, and the differences (AFWA-QNSE) 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 0.35" regardless of initialization time (Fig. 14). Above 0.25" the general trend is a decreasing bias where 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. All SS pair-wise differences are also PS and favor the AFWA configuration. These PS differences are generally noted for the lowest thresholds for forecasts valid at 00 UTC, regardless of the initialization time.

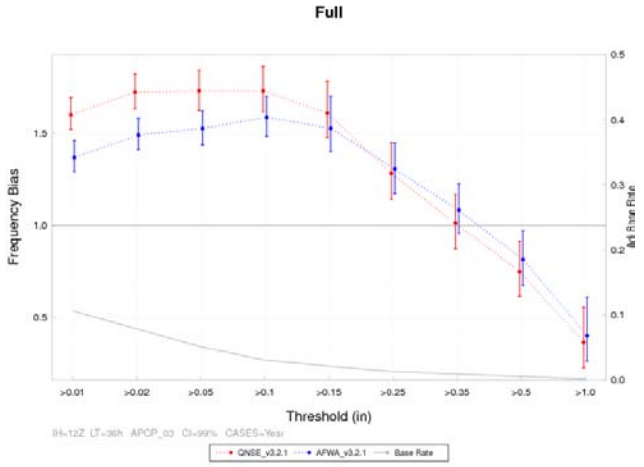

Figure 14. Threshold series plot of 3-hour precipitation accumulation (in) for median frequency bias for the 00 UTC initializations 24-hour lead time only aggregated across the entire year of cases. The AFWA_v3.2.1 configuration is shown in blue and the QNSE_v3.2.1 configuration in red. The vertical 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 Frequency 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 (Fig. 15). Again, there are several SS pair-wise differences noted for thresholds below 1"; however, none are PS.

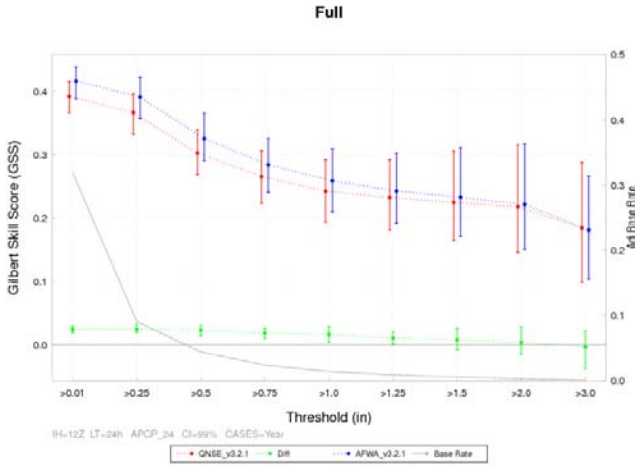

Figure 15. Threshold series plot of 24-hour precipitation accumulation (in) for median GSS for the 12 UTC initializations 24-hour lead time only aggregated across the entire year of cases. The AFWA_v3.2.1 configuration is shown in blue, the QNSE_v3.2.1 configuration in red and the differences (AFWA-QNSE) 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.

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 SS high bias for both configurations (Fig. 16). Once again, when using the more conservative method for assessing SS between the two configurations all favor the AFWA configuration, are PS and occur at the lowest thresholds.

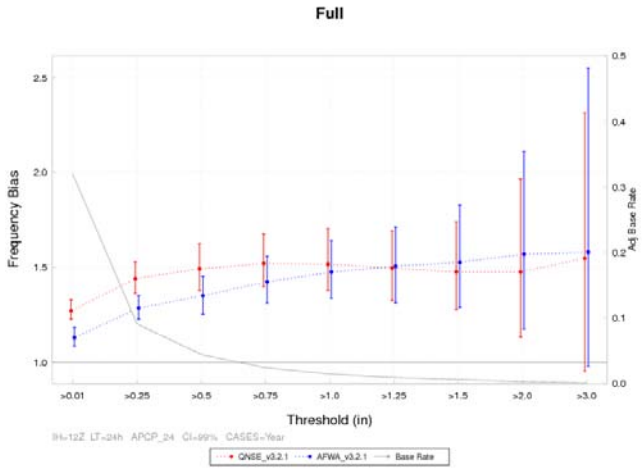

Figure 16. Threshold series plot of 24-hour precipitation accumulation (in) for median frequency bias for the 12 UTC initializations 24-hour lead time only aggregated across the entire year of cases. The AFWA_v3.2.1 configuration is shown in blue and the QNSE_v3.2.1 configuration in red. The vertical 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 1) the new QNSE PBL and surface layer schemes available in WRF, using AFWA's Operational Configuration as a baseline, and 2) the latest WRF version release (v3.2.1), using v3.1.1 as a baseline. Because both configurations for both versions of the code were run for the same cases, pair-wise differences were computed for standard verification metrics between the two configurations and versions, and an assessment of the statistical significance (SS) and practical significance (PS) was included. In general, when examining the AFWA and QNSE configuration using the same WRF version, the AFWA configuration was favored more often. However, for some metrics at certain levels, lead times, or thresholds, QNSE was favored. It should be noted, though, that the relative magnitudes of the SS differences favoring the AFWA configuration are generally larger than those favoring the QNSE configuration leading to a larger number of PS results favoring the AFWA configuration. Please see: http://verif.rap.ucar.edu/eval/afwa_rc_test/ for full details and results of this test and evaluation project.

Because carefully controlled, rigorous testing and evaluation was conducted on these configurations, they have been designated as DTC Reference Configurations (RCs). More details and results on these and other DTC RCs can be found at: http://www.dtcenter.org/config/.

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.

Appendix A: Statistically and Practically Significant Table Statistics for the AFWA_v3.2.1 configuration compared to the QNSE_v3.2.1 configuration.

Table 1. SS (light shading) and PS (dark shading) pair-wise differences for the AFWA and QNSE configurations run with WRF v3.2.1 (where the version highlighted is favored) for the annual aggregation of upper air temperature, dew point temperature and wind BCRMSE and bias by pressure level and forecast lead time for the 00 UTC and 12 UTC initializations combined over the full integration domain.

		Annual											
		Temperature				Dew Point Temperature				Wind			
		f12	f24	f36	f48	f12	f24	f36	f48	f12	f24	f36	f48
BCRMSE	850	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA
	700	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA
	500	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA
	400	--	--	--	AFWA					AFWA	AFWA	AFWA	AFWA
	300	QNSE	QNSE	--	--					AFWA	--	AFWA	AFWA
	200	QNSE	QNSE	QNSE	QNSE					--	--	--	AFWA
	150	AFWA	--	AFWA	AFWA					AFWA	AFWA	AFWA	AFWA
	100	AFWA	AFWA	--	AFWA					AFWA	AFWA	AFWA	AFWA
Bias	850	QNSE	QNSE	--	--	QNSE	QNSE	QNSE	QNSE	AFWA	AFWA	AFWA	AFWA
	700	AFWA	AFWA	AFWA	--	QNSE	QNSE	QNSE	QNSE	QNSE	QNSE	AFWA	AFWA
	500	AFWA	AFWA	AFWA	AFWA	QNSE	QNSE	QNSE	QNSE	QNSE	QNSE	QNSE	QNSE
	400	AFWA	AFWA	--	--					QNSE	QNSE	QNSE	QNSE
	300	--	QNSE	QNSE	QNSE					QNSE	QNSE	QNSE	QNSE
	200	QNSE	QNSE	QNSE	--					QNSE	QNSE	QNSE	QNSE
	150	--	--	AFWA	AFWA					QNSE	QNSE	QNSE	QNSE
	100	QNSE	--	--	--					QNSE	--	--	QNSE

Table 2. SS (light shading) and PS (dark shading) pair-wise differences for the AFWA and QNSE configurations run with WRF v3.2.1 (where the version highlighted is favored) for the annual aggregation of surface temperature, dew point temperature and wind BCRMSE and bias by forecast lead time for the 00 UTC and 12 UTC initializations separately over the full integration domain.

Annual			f03	f06	f09	f12	f15	f18	f21	f24	F27	f30	f33	f36	f39	f42	f45	f48
BRCMSE	00 UTC Initializations	Temperature	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA
		Dew Point Temperature	--	QNSE	QNSE	QNSE	QNSE	--	AFWA	AFWA	QNSE	QNSE	QNSE	QNSE	--	--	AFWA	AFWA
		Wind	QNSE	QNSE	QNSE	--	AFWA	AFWA	AFWA	--	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA
	12 UTC Initializations	Temperature	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA
		Dew Point Temperature	QNSE	AFWA	AFWA	AFWA	--	QNSE	QNSE	--	--	AFWA	AFWA	AFWA	--	--	--	--
		Wind	AFWA	AFWA	AFWA	QNSE	--	--	--	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA
Bias	00 UTC Initializations	Temperature	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA
		Dew Point Temperature	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	QNSE	QNSE	AFWA	AFWA	AFWA
		Wind	QNSE	QNSE	QNSE	QNSE	AFWA	AFWA	AFWA	QNSE	QNSE	QNSE	QNSE	--	AFWA	AFWA	AFWA	QNSE
	12 UTC Initializations	Temperature	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA
		Dew Point Temperature	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	QNSE	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	QNSE
		Wind	AFWA	AFWA	AFWA	QNSE	QNSE	QNSE	QNSE	QNSE	AFWA	AFWA	AFWA	QNSE	QNSE	QNSE	QNSE	--

Appendix B: Statistically and Practically Significant Table Statistics for the AFWA configuration with v3.1.1+ compared to v3.2.1.

Table 1. SS (light shading) and PS (dark shading) pair-wise differences for the AFWA configuration run with WRF v3.1.1+ and v3.2.1 (where the version highlighted is favored) for the annual aggregation of upper air temperature, dew point temperature and wind BCRMSE and bias by pressure level and forecast lead time for the 00 UTC and 12 UTC initializations combined over the full integration domain.

		Annual											
		Temperature				Dew Point Temperature				Wind			
		f12	f24	f36	f48	f12	f24	f36	f48	f12	f24	f36	f48
BCRMSE	850	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	--	V3.1.1	--	V3.1.1
	700	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	--	--	--	--	V3.1.1
	500	--	--	--	V3.2.1	--	--	--	--	V3.2.1	--	--	V3.2.1
	400	--	--	--	--					--	--	--	--
	300	V3.1.1	V3.1.1	--	--					V3.2.1	--	V3.2.1	V3.2.1
	200	V3.1.1	V3.1.1	V3.1.1	V3.1.1					--	--	--	--
	150	--	V3.1.1	V3.1.1	V3.1.1					--	--	--	--
	100	--	--	--	--					V3.2.1	V3.2.1	--	--
Bias	850	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.1.1	--	--	--	V3.2.1	V3.1.1	V3.2.1	V3.1.1
	700	V3.1.1	V3.1.1	V3.1.1	V3.2.1	--	V3.2.1	V3.2.1	--	V3.2.1	V3.2.1	V3.2.1	V3.2.1
	500	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	--	--	--	--
	400	--	V3.1.1	V3.1.1	V3.1.1					V3.2.1	--	--	--
	300	V3.2.1	V3.1.1	V3.1.1	V3.1.1					V3.1.1	V3.1.1	V3.1.1	V3.1.1
	200	V3.2.1	V3.2.1	V3.2.1	V3.2.1					V3.1.1	V3.1.1	--	--
	150	V3.2.1	V3.2.1	V3.2.1	V3.2.1					--	V3.2.1	V3.2.1	V3.2.1
	100	V3.2.1	V3.2.1	V3.2.1	V3.2.1					--	V3.2.1	V3.2.1	V3.2.1

Table 2. SS (light shading) and PS (dark shading) pair-wise differences for the AFWA configuration run with WRF v3.1.1+ and v3.2.1 (where the version highlighted is favored) for the annual aggregation of surface temperature, dew point temperature and wind BCRMSE and bias by forecast lead time for the 00 UTC and 12 UTC initializations separately over the full integration domain.

Annual			f03	f06	f09	f12	f15	f18	f21	f24	F27	f30	f33	f36	f39	f42	f45	f48
BRCMSE	00 UTC Initializations	Temperature	V3.1.1	V3.1.1	V3.1.1	--	--	--	--	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	--	V3.1.1
		Dew Point Temperature	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1				V3.2.1
		Wind	--	--	V3.2.1	V3.2.1	--	V3.1.1	--	--	V3.1.1	V3.1.1	--	V3.1.1	V3.1.1	V3.1.1	V3.1.1	--
	12 UTC Initializations	Temperature	--	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	--	V3.1.1	--	--	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1
		Dew Point Temperature	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	--	--
		Wind	V3.1.1	V3.1.1	--	V3.2.1	V3.1.1	--	--	--	V3.1.1	V3.1.1	V3.1.1	--	V3.1.1	V3.1.1	V3.1.1	V3.1.1
Bias	00 UTC Initializations	Temperature	V3.1.1	V3.2.1	V3.2.1	V3.2.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.2.1	V3.2.1	V3.2.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1
		Dew Point Temperature	V3.1.1	V3.1.1	--	--	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.1.1	V3.1.1	V3.1.1
		Wind	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.1.1	V3.1.1	V3.1.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.1.1	V3.1.1	V3.1.1	V3.2.1
	12 UTC Initializations	Temperature	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.1.1	V3.2.1	V3.2.1	V3.1.1
		Dew Point Temperature	V3.2.1	V3.1.1	V3.1.1	V3.1.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.1.1	V3.1.1	V3.1.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1
		Wind	V3.1.1	V3.1.1	V3.1.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.1.1	V3.1.1	V3.1.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1	V3.2.1