The Developmental Testbed Center
WRF-RR Vertical Level (WRFRR-VL) Sensitivity Test
Final Report

Point of Contact: Jamie Wolff
March 31, 2010

Executive Summary

To assess the sensitivity of forecast verification statistics to the vertical placement of near-surface sigma levels in the Weather Research and Forecasting (WRF) model, the Developmental Testbed Center (DTC) ran a carefully controlled test and evaluation of forecasts generated with the Advanced Research WRF (ARW) core using two model configurations that were identical except for the distribution of the nine lowest sigma levels. This report focuses on the differences between the standard verification measures for the two configurations, including an assessment of the statistical significance. The following points summarize the statistically significant (SS) differences as seen in the verification statistics between the high resolution configuration (ARW1) and the default resolution (ARW2):

- The only SS pair-wise differences found in the vertical profile comparison between ARW1 and ARW2 were generally seen at 850 hPa level for temperature bias where ARW1 more frequently had a smaller cold bias, and for wind speed bias where ARW2 had a smaller high bias.
- The SS pair-wise differences for the surface temperature forecasts initialized at both 00 and 12 UTC indicate that the higher resolution configuration (ARW1), in general, has SS lower BC-RMSE values during the overnight hours and SS higher values during the daytime hours. The same general trend is noted for surface relative humidity BC-RMSE and bias, however, for BC-RMSE values of surface vector winds all SS pair-wise differences favor the ARW1 configuration.
- All of the SS pair-wise differences for surface temperature and wind speed bias favor the ARW1 configuration. This indicates that the surface temperature from the ARW1 configuration has a smaller cold bias during the day and smaller warm bias overnight, while for surface wind speed the ARW1 configuration has a consistently smaller high bias.
- No SS differences between the two configurations are seen in the ETS of 3-hourly or daily (24-hour) precipitation accumulation. For the 3-hourly accumulation bias, the ARW2 configuration has a SS lower high bias for forecasts valid at 00 UTC (regardless of initialization time/lead time) at several thresholds. For the daily precipitation accumulation bias, all SS pair-wise differences also favor the ARW2 configuration.

1. Introduction

To assess the sensitivity of forecast verification statistics for atmospheric variables near the land surface (2-m temp and relative humidity, 10-m wind, as well as precipitation) and on land-surface variables (particularly snow water equivalent and snowmelt) to the vertical placement of near-surface sigma levels in the Weather Research and Forecasting (WRF) model, the Developmental Testbed Center (DTC; Bernardet et al. 2008) ran a carefully controlled test and evaluation of forecasts generated with the Advanced Research WRF (ARW) dynamic core (Skamarock et. al, 2008) using two model configurations that were identical except for the distribution of the nine lowest sigma levels. This task was in support of the implementation of the WRF Rapid Refresh (WRF-RR) at the National Centers for Environmental Prediction (NCEP).

2. Experiment Design

The components of the end-to-end forecast system used in the WRFRR-VL Test included the officially released v3.0 codes for the WRF Preprocessing System (WPS), WRF-ARW model, and WRF Post Processor (WPP), with minor bug fixes incorporated in WPP before the comprehensive test was run. Graphics were generated using the NCAR Command Language (NCL). The NCEP Verification System was utilized to verify the post-processed forecasts and confidence intervals were computed using routines developed by the DTC in the statistical programming language, R. In addition, the full data set was archived and made available for dissemination.

2.1 Forecast Periods

Retrospective forecasts were run from 25 March – 25 April 2006. The forecasts were initialized at 00 and 12 UTC every day and run out to 24 hours with output files every three hours. Forecasts for which the end-to-end system was not able to run to completion are summarized in the tables below. A total of 8 out of the possible 62 cases did not complete from initialization to verification. Of those incomplete cases, 5 failed due to missing or corrupt model input data and 3 failed due to missing verification observations; none were due to model crashes.

Missing forecasts:

Affected case	Missing data	Reason
2006032500	WRF output	Corrupt RUC input data
2006040200	WRF output	Corrupt RUC input data
2006041300	WRF output	Missing Eta input data
2006041712	WRF output	Missing RUC input data
2006041800	WRF output	Missing RUC input data

Missing verification:

Affected case	Missing data	Reason
2006032912	Incomplete 24-h QPF verification	Missing RFC analysis
2006042300	Incomplete sfc/upa verification	Missing RUC Prepbufr files
2006042312	Incomplete 24-h QPF and sfc/upa verification	Missing RFC analysis and RUC Prepbufr files

2.2 Initial and Boundary Conditions

The WRFRR-VL Test utilized initial conditions (ICs) from the Rapid Update Cycle (RUC13) model and lateral boundary conditions (LBCs) from the North American Mesoscale Model (NAM212). For this retrospective period, the forecast component of the NAM was the Eta model. This mixture of input data was chosen because RUC13 grids are not available out to 24 hours for the retrospective time period. The NAM grids from the prior 06 UTC and 18 UTC initializations were used to produce LBCs for the 12 UTC and 00 UTC forecasts, respectively, in order to mimic the RUC13 operational setup. In order to use these two sources of input data, the ***real*** (WRF Initialization) program was run twice; first with the NAM input to generate the required LBC file and, second with the RUC input to generate the required IC file.

In addition, NCEP's daily, real-time sea surface temperature (SST) product was used to initialize the SST field for the forecasts. This SST product is produced on a one-twelfth degree latitude-longitude grid using a two-dimensional variational interpolation analysis of the most recent 24-hours ship and buoy data, satellite-retrieved SST data, and SSTs derived from satellite-observed sea-ice coverage. Although the end-to-end system for this testing did not include a data assimilation component, the RUC13 cloud fields were included in the initial conditions, so the runs were not truly cold start forecasts in the sense of starting with no clouds; there was, however, no attempt to create an initial divergent wind component that is consistent with the initial clouds.

In addition, the time-invariant components of the lower boundary conditions (topography, soil and vegetation type etc.) were generated through the ***geogrid*** program of WPS.

2.3 Model Configuration Specifics

2.3.1 Model Configuration

The ARW dynamical core was used for this test. The timesteps used were a long timestep of 72 s and an acoustic timestep of 18 s. Calls to the boundary layer, microphysics and cumulus parameterization were made every time step, whereas calls to radiation were made every 30 minutes. Output files were produced every three hours.

The ARW solver offers a number of run-time options for the numerics, as well as various filter and damping options (Skamarock et al. 2008). For this test, the ARW was configured to use the following numeric options: 3rd-order Runge-Kutta time integration, 5th-order horizontal momentum and scalar advection, and 3rd-order vertical momentum and scalar advection. In addition, the following filter/damping options were utilized: three-dimensional divergence damping (coefficient 0.1), external mode filter (coefficient 0.01), off-center integration of vertical momentum and geopotential equations (coefficient 0.1), vertical-velocity damping (only applied to those grid points where there are locally strong updraft cores), and a 5-km-deep diffusive damping layer at the top of the domain (coefficient 0.02).

Relevant sections of the ***namelist.input*** file can be found in Appendix A.

2.3.2 Domain Configuration

The roughly 13-km domain used for this test (Figure 1) was selected such that it fits within the RUC13 domain. It has 400 x 304 gridpoints, for a total of 121,600 horizontal gridpoints. The Lambert-Conformal map projection was used and the model was configured to have 50 vertical levels (51 sigma entries). The sigma level distribution for both configurations is shown in Figure 2 and the delta sigma values in Figure 3. Effort was made to ensure a smooth transition between the levels for while increasing the resolution for the lowest nine sigma levels.

2.3.3 Physics Suite

The physics suite used for this test is described in the table below.

Physics suite used for WRFRR-VL test:

Microphysics	Thompson
Surface Layer	Janjic
Planetary Boundary Layer	Mellor-Yamada-Janjic
Convection	Grell-Devenyi ensemble
Land-Surface Model	RUC
Short/Long wave Radiation	Dudhia/RRTM

2.4 Post-processing

The WPP (Chuang et al. 2004) was used to destagger the forecasts, to generate derived meteorological variables, including mean sea level pressure, and to vertically interpolate fields to isobaric levels. The post-processed files include two-dimensional fields, three-dimensional fields on constant pressure levels (which are required by the plotting and verification programs), and three-dimensional post-processed fields on model native vertical coordinates (used to plot the skew-T soundings).

3. Model Verification

Model verification partial sums (aggregations by geographical region using the mean) were generated using the NCEP Verification System (Chuang et al. 2004), which is comprised of the Surface and Upper Air Verification System and the Quantitative Precipitation Forecast (QPF) Verification System. Objective model verification statistics were then computed from these partial sums using the statistical programming language, R. For precipitation, the area-averaged aggregations were done by summing up the contingency tables for all cases run and computing the scores based on the cumulated table. For all the other variables, the median of the area averaged statistic (which is considered a robust statistical measure) was used.

Several domains were considered for the verification of surface and upper air, as well as precipitation. Area-average results were computed using the NCEP Verification System over the entire CONUS (G164), CONUS-West (G165), and CONUS-East (G166) domains (Figure 4). The surface and upper air components of the NCEP Verification System were also configured to compute area-averages for the 14 regional domains shown in Figure 5. Only results from CONUS, CONUS-West and CONUS-East are presented in this report. In addition to the regional area stratification, the verification statistics were also stratified by lead time and vertical level for the 00 UTC and 12 UTC initialization hours combined, except for surface verification where forecasts were also stratified by initialization hour (00 and 12 UTC) in order to preserve any diurnal signal.

Confidence intervals (CIs), at the 99% level, were applied to each of the variables using the appropriate statistical method. Since verification statistics are only computed for cases that ran to completion for both configurations of the model, pair-wise differences between the verification statistics of the two configurations are computed and also presented. 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.

3.1 Temperature, relative humidity and winds

Forecasts for temperature, relative humidity and winds at the surface and upper-air were bilinearly interpolated to the location of the observations (METARs and RAOBS) in the RUC Prepbufr files using the NCEP Surface and Upper Air Verification System and a grid-to-point comparison was performed. Objective model verification statistics were generated for surface and upper-air temperature, relative humidity and winds. At the surface, 2-m temperature and relative humidity and 10-m winds are compared to METAR observations and verification statistics computed at 3-hour intervals. Because the values of the shelter-level variables are not appropriate to verify at the model initial time, surface verification results start at the 3-hour lead time. For upper-air, verification statistics were computed at mandatory levels using radiosonde observations and computed at 12-hour intervals starting at the initialization time. Because of known errors associated with radiosonde humidity measurements at high altitudes, the analysis of the upper air relative humidity verification focuses on levels at and below 500 hPa. The verification measures included in this report are the biased-corrected root mean square error (BC-RMSE) and the mean error (bias), computed separately for each observational type. Correction for autocorrelation was computed and CIs, computed assuming a normal distribution, were applied to each of these measures.

3.2 Precipitation

For the precipitation verification, the NCEP Quantitative Precipitation Forecast (QPF) Verification System was used to perform a grid-to-grid comparison in which the forecasts and the precipitation analyses were first interpolated to Grid 218 (12-km grid) and then evaluated. Accumulation periods of 3 hours and 24 hours were examined. The observational datasets for the 3-hourly accumulations are the NCEP Stage II data, while the NCEP/CPC daily gauge analysis was used for the 24-hour accumulation (valid at 12 UTC). Because forecasts were only generated out to 24 hours, verification for the 24-hour accumulation was only possible for the forecasts initialized at 12 UTC. The verification measures computed for these fields were the equitable threat score (ETS) and the frequency bias. A resampling (bootstrap) technique was used to compute the applied CIs.

4. Verification Results

The model configuration with higher vertical resolution near the surface will be referred to as ARW1, while the configuration with a sigma level distribution similar to the default values distributed in the official WRF v3.0 tar file will be referred to as ARW2. Differences are computed between the two configurations by subtracting ARW2 from ARW1. (For more information on how each of these statistics is computed, please see Appendix B). A breakdown of the configuration with SS better performance by variable, statistic, initialization hour, forecast lead time, and level is summarized in Tables 1-8, where the favored configuration with the SS better score is highlighted.

4.1 Upper Air

4.1.1 Temperature BC-RMSE and bias

As expected, the BC-RMSE values increased from the initial time to the 24-hour lead time at all vertical levels examined (Figure 6). A similar vertical distribution is noted for the CONUS and

sub-CONUS domains at the 12- and 24-hour lead times, with a minimum value occurring around 500 hPa and a maximum value at the upper most levels. No SS pair-wise differences are seen for any domain, level, or lead time (Table 1).

When examining the bias for the two configurations over all the domains, there is a cold bias maximum for the 12- and 24-hour lead times at 150 hPa and a secondary maximum at 700 hPa (Figure 7). Conversely, a warm bias maximum is noted at 200 and 100 hPa. The magnitudes of these extrema are smallest for the initial time and increase with lead time. For the CONUS domain, the bias for the 12- and 24-hour lead times do not differ significantly from a non-biased at 400 and 300 hPa. The CONUS-West domain has SS colder bias when compared to CONUS-East for the 12- and 24-hour lead times for the lowest levels. For the 12-hour lead time the only SS pair-wise difference noted occurs for the CONUS-East domain and favors the ARW2 configuration, while for the 24-hour lead time SS pair-wise differences are noted for the CONUS domain at 850 hPa and the CONUS-East domain at the 850 and 700 hPa levels, all favoring the ARW1 configuration (Table 1).

4.1.2 Relative Humidity BC-RMSE and bias

An increase in BC-RMSE values with lead time is noted for relative humidity; however, no SS pair-wise differences are noted for the CONUS domain (Figure 8) anywhere in the vertical profile. This result holds for all forecast lead times, even when further stratified by domain (not shown).

The magnitudes of the bias for both configurations at each lead time are very similar for all levels examined, i.e. there is not a substantial increase with lead time (Figure 9). Because the CIs encompass zero for the 12- and 24-hour lead times at and below 700 hPa, it is not possible to say these differ significantly from a non-biased forecast. Above this level for all lead times (including the 00-hour), a moist bias for both configurations is noted. There are no SS pair-wise differences for relative humidity bias at any level, forecast lead time, or domain (Table 2).

4.1.3 Winds BC-RMSE and bias

For the vector wind comparison on the CONUS domain, the BC-RMSE values increase more at mid- to upper-levels than lower-levels as the lead time increases (Figure 10). This variable shows very little difference between the two configurations across the lead times, levels, and domain stratifications examined, with no SS pair-wise differences revealed (Table 3).

When examining the wind speed bias, at the initial time the winds are too light for most levels. This result holds for the 12- and 24-hour lead times at 500, 250, 200 and 150 hPa, while the winds are too strong at 850 hPa (Figure 11). All other levels (700, 400, 300, and 150 hPa) encompass zero (no bias) indicating a smaller wind speed bias for the 12- and 24-hour lead times than the initial time. When comparing the two configurations, the pair-wise differences are SS for all domains (sub-domains not shown) at 850 hPa for several lead times (Table 3). The ARW2 configuration has a smaller high bias, and is favored.

4.2 Surface

4.2.1 Temperature BC-RMSE and bias

For surface temperature, in addition to domain, the verification results are stratified by initialization time in order to separate out the diurnal effects. Both configurations have a similar distribution of BC-RMSE values for both the 00 and 12 UTC initializations (Figure 12). There is only a small growth in the error values as the forecast lead time increases to 24 hours. For example, when comparing the 6-hour lead time from the 12 UTC initializations with the 18-hour

lead time from the 00 UTC initializations (both valid at 18 UTC), the median value increases for the longer lead time, however, the increase is not SS because the range of values encompassed by the CIs overlap. Regardless of the domain examined, for the 00 UTC initializations, all SS pair-wise differences favor the ARW1 configuration for the first 12 hours of the forecast (Table 4). Conversely, for the 15- to 21-hour lead times the SS pair-wise differences generally favor the ARW2 configuration. The opposite trends are noted for the 12 UTC initializations, where the ARW2 configuration is favored for all SS pair-wise differences between the 6- and 12-hour lead times, while the ARW1 configuration is generally favored for forecast hours beyond 12 hours. This result is directly related to the diurnal cycle and indicates that the higher resolution configuration (ARW1) generally has lower BC-RMSE values during the overnight hours and higher values during the daytime hours, when compared to the lower resolution configuration (ARW2).

For both the 00 and 12 UTC initializations, during the overnight hours (forecasts valid between 03 - 12 UTC), all three domains have a warm surface temperature bias (Figure 13). The CONUS-West domain surface temperature forecasts valid between 15 and 00 UTC (daytime hours) reveal a cold bias, while for the CONUS and CONUS-East domains the CIs generally encompass zero for forecasts valid between 18 and 00 UTC, thus, it is not possible to say the forecasts are significantly different from a non-biased forecast. All SS pair-wise differences favor the ARW1 configuration, indicating that, in general, the ARW1 configuration has a smaller cold bias during the day and a smaller warm bias overnight (Table 4).

4.2.2 Relative Humidity BC-RMSE and bias

The distribution of surface relative humidity BC-RMSE values for all domains (CONUS only shown) are very similar (Figure 14; top), and because of the direct relationship with temperature, are generally consistent with the distribution seen for surface temperature errors described above. For the SS pair-wise differences, it is again noted that the ARW1 configuration is generally favored during the overnight hours while the ARW2 configuration is generally favored during the daytime. However, overall fewer SS pair-wise differences are noted for surface relative humidity as compared to surface temperature (Table 5).

The surface relative humidity bias from the 00 UTC initializations indicate a SS bias for the 12- through 18-hour lead times (dry bias at 12 UTC and moist bias for 15 and 18 UTC), while for all other forecast lead times the CIs encompass zero (Figure 14; bottom). For the 12 UTC initializations (not shown), the CIs for both configurations at all lead times encompass zero, signifying neither configuration is significantly different from a non-biased forecast. Even so, SS pair-wise differences favor the ARW1 configuration for both the 00 and 12 UTC initializations during the overnight hours (Table 5). Though there are fewer SS pair-wise differences noted during the daytime hours, they generally favor the ARW2 configuration.

4.2.3 Wind BC-RMSE and bias

For the surface vector wind BC-RMSE values the CONUS-WEST domain has consistently higher error values than CONUS-East (Figure 15; top row). The general characteristic previously noted of small error growth with increasing lead time is again demonstrated for this parameter shown by a small difference in BC-RMSE values when comparing a 12-hour forecast from a 12 UTC initialization and a 24-hour forecast from a 00 UTC initialization. All SS pair-wise differences noted for surface vector wind errors favor the ARW1 configuration, with the CONUS-West domain indicating the largest number of differences between the two configurations (Table 6).

A consistent high bias for both configurations and all domains is noted for surface wind speed, and, overall, bias values are SS higher for the CONUS-East domain compared to the CONUS-West domain (Figure 15; bottom). SS pair-wise differences for wind speed bias occur at most

lead times, for both initialization times, for each domain, all of which favor the ARW1 configuration (Table 6).

4.2.4 3-hourly Precipitation ETS and bias

In general, for 3-hourly precipitation accumulation, regardless of the stratification (configuration, initialization time, domain, forecast lead time (12 and 24 only, examined)), the equitable threat score decreases from a maximum at the lowest (0.05") threshold to a minimum at the highest (0.5") threshold (Figure 16). There are no SS pair-wise differences for any of the above stratifications (Table 7).

There is a high bias noted in the 3-hourly precipitation accumulation for both configurations initialized at 00 UTC for the 12- and 24-hour lead times for thresholds at and below 0.15" (Figure 16; bottom row). Above that threshold the confidence intervals encompass one and thus are not significantly different from an unbiased forecast. All SS pair-wise differences favor the ARW2 configuration (Table 7). These SS pair-wise differences are only seen for forecasts valid at 00 UTC (i.e. 12-hour forecasts from the 12 UTC initializations and 24-hour forecasts from the 00 UTC initializations) and span across many of the thresholds.

4.2.5 Daily Precipitation ETS and bias

Due to the fact that the cases were only run out 24 hours and the daily precipitation observations are taken at 12 UTC, the 24-hour precipitation accumulation could only be examined for the 12 UTC initializations. As with the 3-hourly precipitation accumulation, the ETS values ranged from a high at the 0.05" threshold to a low at the 2" threshold for all domains (Figure 17). There are, again, no SS pair-wise differences for the 24-hour precipitation accumulation ETS value, regardless of domain (Table 8).

In general, the bias for the daily precipitation is not significantly different from one for both configurations at the highest thresholds (Figure 18). As was seen for the 3-hour accumulations, all of the SS pair-wise differences noted for each of the domains favor the ARW2 configuration, which has a lower high bias for several of the thresholds (Table 8).

5. Summary

Two WRF-ARW configurations were comprehensively tested and evaluated to assess the sensitivity of forecast verification statistics for atmospheric variables near the land surface to the vertical placement of near-surface sigma levels. The first configuration (ARW1) was a high-resolution configuration near the surface while and the second (ARW2) used a similar distribution to the default resolution distributed in the WRF v3.0 tar file. 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 statistical significance was included. In general, for the vertical profiles, the two configurations were not significantly different, with only a few exceptions (850 hPa bias for temperature and wind speed). For surface related fields, the BC-RMSE values of temperature and relative humidity indicated that the ARW1 configuration was favored during the overnight hours, while the ARW2 configuration was favored during the day time. This general trend holds for surface relative humidity bias, but for surface temperature and wind BC-RMSE and bias all SS pair-wise differences favored the ARW1 configuration. Finally, for ETS of both the 3-hourly and daily precipitation accumulations there were no SS pair-wise differences, while all SS pair-wise differences for bias favored the ARW2 configuration.

6. References

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

Chuang, H.-Y., G. DiMego, M. Baldwin, and WRF DTC Team, 2004: NCEP's WRF post-processor and verification systems. 5th WRF/14th MM5 Users' Workshop, 22-25 June 2004, Boulder, CO.

Nance L., 2006. Weather Research and Forecasting Core Test – DTC Report (available from http://ruc.fsl.noaa.gov/coretest2/DTC_report.pdf).

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.

Table 1. SS differences for upper air temperature BCRMSE and bias by pressure level, domain, and forecast lead time for the 00 UTC and 12 UTC initializations combined.

		CONUS			CONUS-East			CONUS-West		
		f00	f12	f24	f00	f12	f24	f00	f12	f24
BCRMSE	850 hPa	--	--	--	--	--	--	--	--	--
	700 hPa	--	--	--	--	--	--	--	--	--
	500 hPa	--	--	--	--	--	--	--	--	--
	400 hPa	--	--	--	--	--	--	--	--	--
	300 hPa	--	--	--	--	--	--	--	--	--
	200 hPa	--	--	--	--	--	--	--	--	--
	150 hPa	--	--	--	--	--	--	--	--	--
	100 hPa	--	--	--	--	--	--	--	--	--
Bias	850 hPa	--	--	ARW1	--	ARW2	ARW1	--	--	--
	700 hPa	--	--	--	--	--	ARW1	--	--	--
	500 hPa	--	--	--	--	--	--	--	--	--
	400 hPa	--	--	--	--	--	--	--	--	--
	300 hPa	--	--	--	--	--	--	--	--	--
	200 hPa	--	--	--	--	--	--	--	--	--
	150 hPa	--	--	--	--	--	--	--	--	--
	100 hPa	--	--	--	--	--	--	--	--	--

Table 2. SS differences for upper air relative humidity BCRMSE and bias by pressure level, domain, and forecast lead time for the 00 UTC and 12 UTC initializations combined.

		CONUS			CONUS-East			CONUS-West		
		f00	f12	f24	f00	f12	f24	f00	f12	f24
BCRMSE	850 hPa	--	--	--	--	--	--	--	--	--
	700 hPa	--	--	--	--	--	--	--	--	--
	500 hPa	--	--	--	--	--	--	--	--	--
Bias	850 hPa	--	--	--	--	--	--	--	--	--
	700 hPa	--	--	--	--	--	--	--	--	--
	500 hPa	--	--	--	--	--	--	--	--	--

Table 3. SS differences for upper air wind BCRMSE and bias by pressure level, domain, and forecast lead time for the 00 UTC and 12 UTC initializations combined.

		CONUS			CONUS-East			CONUS-West		
		f00	f12	f24	f00	f12	f24	f00	f12	f24
BCRMSE	850 hPa	--	--	--	--	--	--	--	--	--
	700 hPa	--	--	--	--	--	--	--	--	--
	500 hPa	--	--	--	--	--	--	--	--	--
	400 hPa	--	--	--	--	--	--	--	--	--
	300 hPa	--	--	--	--	--	--	--	--	--
	200 hPa	--	--	--	--	--	--	--	--	--
	150 hPa	--	--	--	--	--	--	--	--	--
	100 hPa	--	--	--	--	--	--	--	--	--
Bias	850 hPa	--	ARW2	ARW2	--	--	ARW2	ARW2	--	ARW2
	700 hPa	--	--	--	--	--	--	--	--	--
	500 hPa	--	--	--	--	--	--	--	--	--
	400 hPa	--	--	--	--	--	--	--	--	--
	300 hPa	--	--	--	--	--	--	--	--	--
	200 hPa	--	--	--	--	--	--	--	--	--
	150 hPa	--	--	--	--	--	--	--	--	--
	100 hPa	--	--	--	--	--	--	--	--	--

Table 4. SS differences for surface temperature BCRMSE and bias by domain and forecast lead time for the 00 UTC and 12 UTC initializations separately.

			f03	f06	f09	f12	f15	f18	f21	f24
BCRMSE	00 UTC Inits	CONUS	ARW1	ARW1	ARW1	ARW1	ARW2	ARW2	ARW2	--
		CONUS-East	ARW1	ARW1	ARW1	ARW1	--	ARW2	--	--
		CONUS-West	--	ARW1	ARW1	ARW1	--	ARW2	ARW2	ARW1
	12 UTC Inits	CONUS	--	ARW2	ARW2	--	ARW1	ARW1	ARW1	ARW1
		CONUS-East	--	--	ARW2	--	ARW1	ARW1	ARW1	ARW1
		CONUS-West	--	ARW2	ARW2	ARW2	--	ARW2	ARW1	ARW2
Bias	00 UTC Inits	CONUS	--	ARW1	ARW1	ARW1	ARW1	ARW1	ARW1	--
		CONUS-East	--	--	ARW1	--	ARW1	ARW1	ARW1	--
		CONUS-West	ARW1	--	--	ARW1	ARW1	ARW1	ARW1	ARW1
	12 UTC Inits	CONUS	ARW1	ARW1	ARW1	--	ARW1	ARW1	ARW1	ARW1
		CONUS-East	ARW1	ARW1	ARW1	--	ARW1	ARW1	ARW1	ARW1
		CONUS-West	ARW1	--	--	ARW1	ARW1	ARW1	ARW1	ARW1

Table 5. SS differences for surface relative humidity BCRMSE and bias by domain and forecast lead time for the 00 UTC and 12 UTC initializations separately.

			f03	f06	f09	f12	f15	f18	f21	f24
BCRMSE	00 UTC Inits	CONUS	--	--	ARW1	ARW1	--	--	ARW2	--
		CONUS-East	--	--	ARW1	ARW1	ARW2	ARW2	ARW2	ARW2
		CONUS-West	--	ARW1	ARW1	ARW1	--	--	ARW2	ARW2
	12 UTC Inits	CONUS	--	ARW2	ARW2	ARW2	--	--	ARW1	ARW1
		CONUS-East	ARW2	--	ARW2	ARW2	--	--	ARW1	--
		CONUS-West	--	ARW2	ARW2	ARW2	--	--	ARW1	ARW1
Bias	00 UTC Inits	CONUS	ARW1	ARW1	ARW1	ARW1	--	--	--	ARW2
		CONUS-East	ARW1	ARW1	ARW1	ARW1	ARW1	--	--	ARW2
		CONUS-West	ARW1	--	--	ARW1	ARW1	ARW1	ARW2	ARW2
	12 UTC Inits	CONUS	ARW1	--	--	ARW2	ARW1	ARW1	ARW1	ARW1
		CONUS-East	ARW1	--	--	ARW2	ARW1	ARW1	ARW1	ARW1
		CONUS-West	ARW2	--	--	ARW2	ARW2	ARW1	ARW1	ARW1

Table 6. SS differences for surface wind BCRMSE and bias by domain and forecast lead time for the 00 UTC and 12 UTC initializations separately.

			f03	f06	f09	f12	f15	f18	f21	f24
BCRMSE	00 UTC Inits	CONUS	--	--	--	--	--	--	--	--
		CONUS-East	--	--	--	--	--	--	--	ARW1
		CONUS-West	--	ARW1	ARW1	ARW1	--	ARW1	ARW1	ARW1
	12 UTC Inits	CONUS	--	--	--	ARW1	--	--	--	--
		CONUS-East	--	--	--	ARW1	--	--	--	--
		CONUS-West	--	ARW1	ARW1	ARW1	--	ARW1	ARW1	ARW1
Bias	00 UTC Inits	CONUS	ARW1	ARW1	ARW1	ARW1	--	ARW1	ARW1	ARW1
		CONUS-East	ARW1	ARW1	ARW1	ARW1	--	ARW1	ARW1	ARW1
		CONUS-West	ARW1	--	--	ARW1	ARW1	ARW1	ARW1	ARW1
	12 UTC Inits	CONUS	--	ARW1	ARW1	ARW1	ARW1	ARW1	ARW1	ARW1
		CONUS-East	--	ARW1	ARW1	ARW1	ARW1	ARW1	ARW1	ARW1
		CONUS-West	ARW1	--	--	ARW1	ARW1	ARW1	ARW1	ARW1

Table 7. SS differences for 3-hour QPF ETS and frequency bias by domain, forecast lead time, and threshold for the 00 UTC and 12 UTC initializations separately.

			00 UTC Initializations								12 UTC Initializations							
			>0.01”	>0.02”	>0.05”	>0.1”	>0.15”	>0.25”	>0.35”	>0.5”	>0.01”	>0.02”	>0.05”	>0.1”	>0.15”	>0.25”	>0.35”	>0.5”
ETS	CONUS	f12	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		f24	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	CONUS-East	f12	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		f24	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	CONUS-West	f12	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		f24	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Frequency Bias	CONUS	f12	--	--	--	--	--	--	--	--	ARW2	ARW2	ARW2	ARW2	--	--	--	ARW2
		f24	ARW2	ARW2	ARW2	ARW2	ARW2	ARW2	ARW2	--	--	--	--	--	--	--	--	--
	CONUS-East	f12	--	--	--	--	--	--	--	--	ARW2	ARW2	ARW2	ARW2	ARW2	--	--	ARW2
		f24	ARW2	ARW2	ARW2	ARW2	ARW2	ARW2	ARW2	--	--	--	--	--	--	--	--	--
	CONUS-West	f12	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		f24	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Table 8. SS differences for 24-hour QPF ETS and frequency bias by domain and threshold for the 24-hour lead time of the 12 UTC initializations only.

			12 UTC Initializations							
			>0.01"	>0.1"	>0.25"	>0.5"	>0.75"	>1.0"	>1.5"	>2.0"
ETS	CONUS	f24	--	--	--	--	--	--	--	--
	CONUS-East	f24	--	--	--	--	--	--	--	--
	CONUS-West	f24	--	--	--	--	--	--	--	--
Frequency Bias	CONUS	f24	--	--	ARW2	ARW2	ARW2	ARW2	--	--
	CONUS-East	f24	--	--	--	ARW2	ARW2	--	--	--
	CONUS-West	f24	--	ARW2	ARW2	--	--	--	--	--

Figure 1. Map showing the boundary of the WRF computational domain (dashed line).

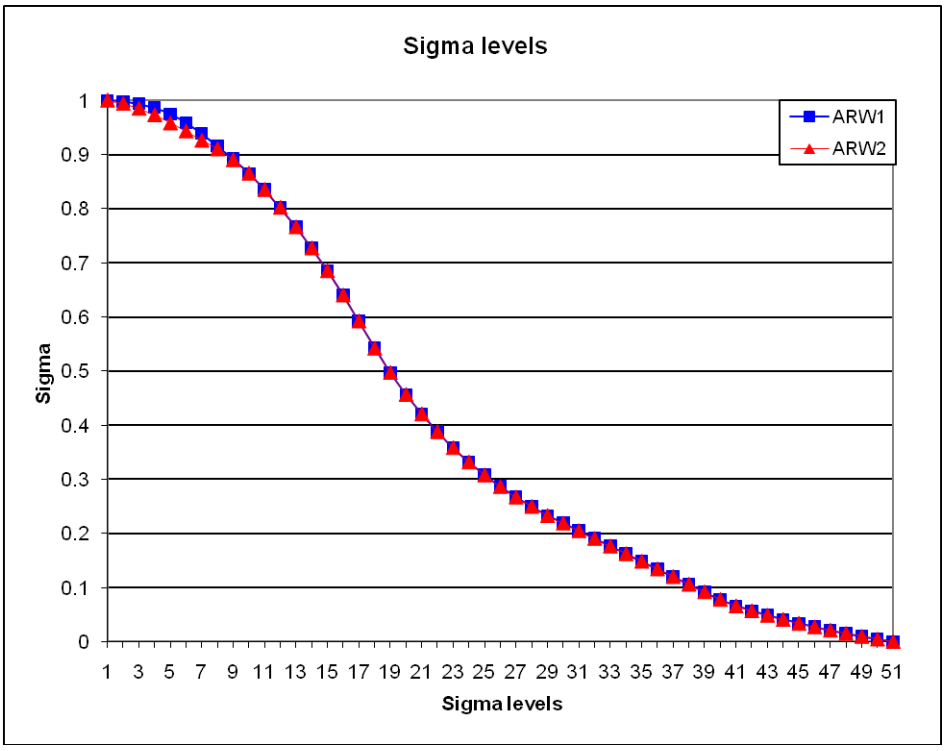

Figure 2. Sigma levels for the ARW1 (blue) and ARW2 (red) configurations.

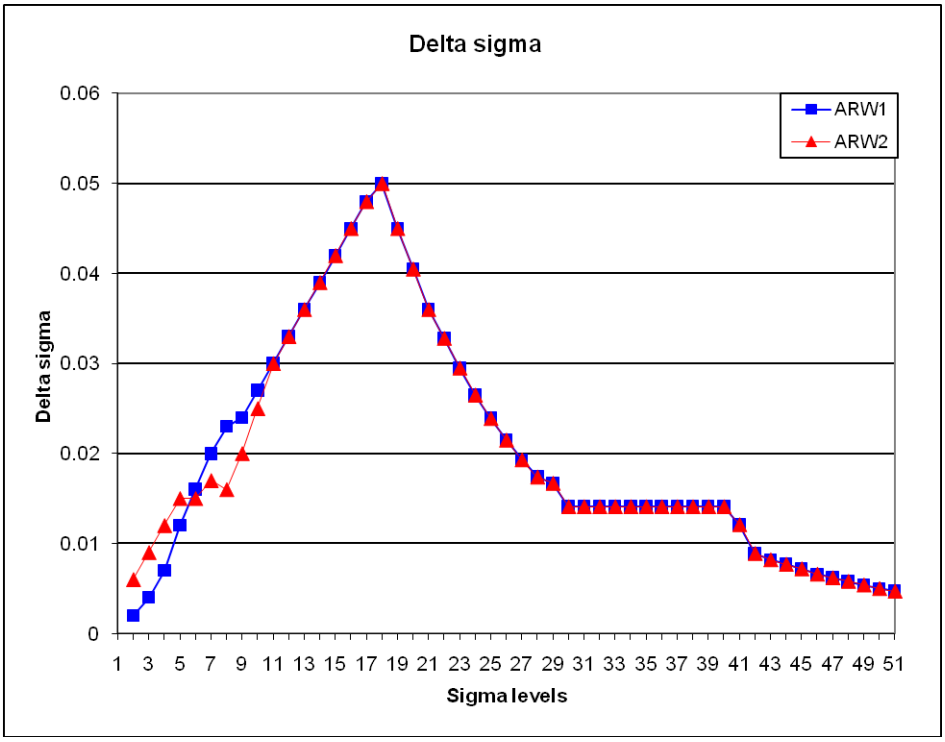

Figure 3. Differences between the sigma levels for the ARW1 (blue) and ARW2 (red) configurations.

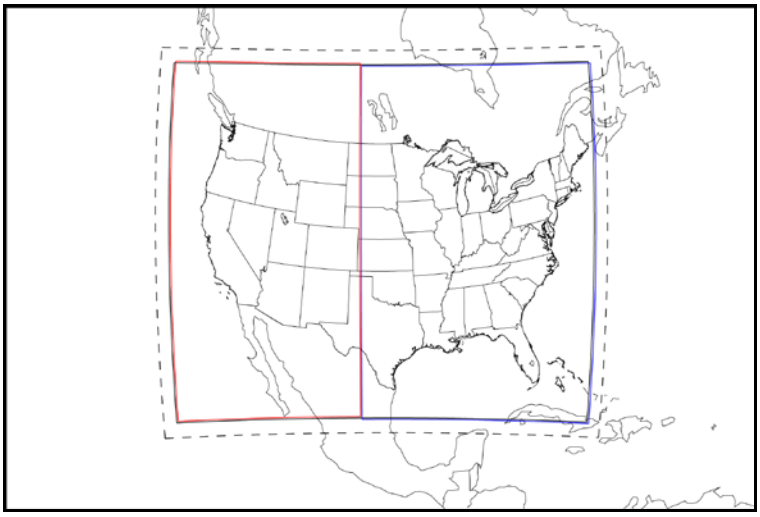

Figure 4. Map showing the boundaries of the verification domains: CONUS (solid black), CONUS-West (solid red), and CONUS-East (solid blue). The WRF computational domain (dashed line) is also shown for reference.

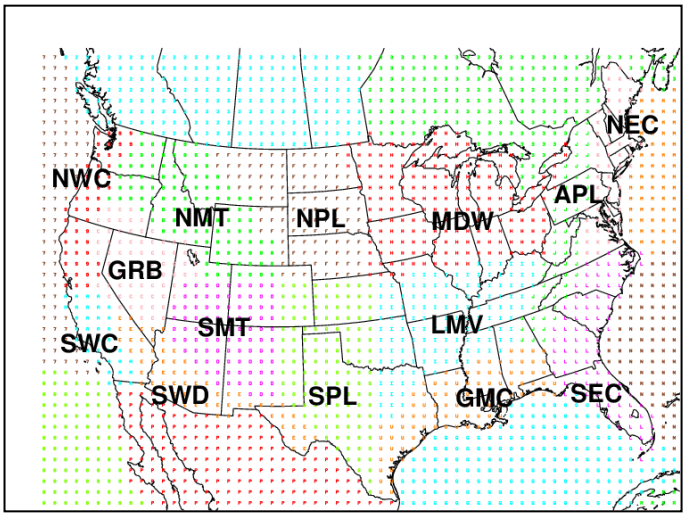

Figure 5. Map showing the locations of the 14 regional verification domains.

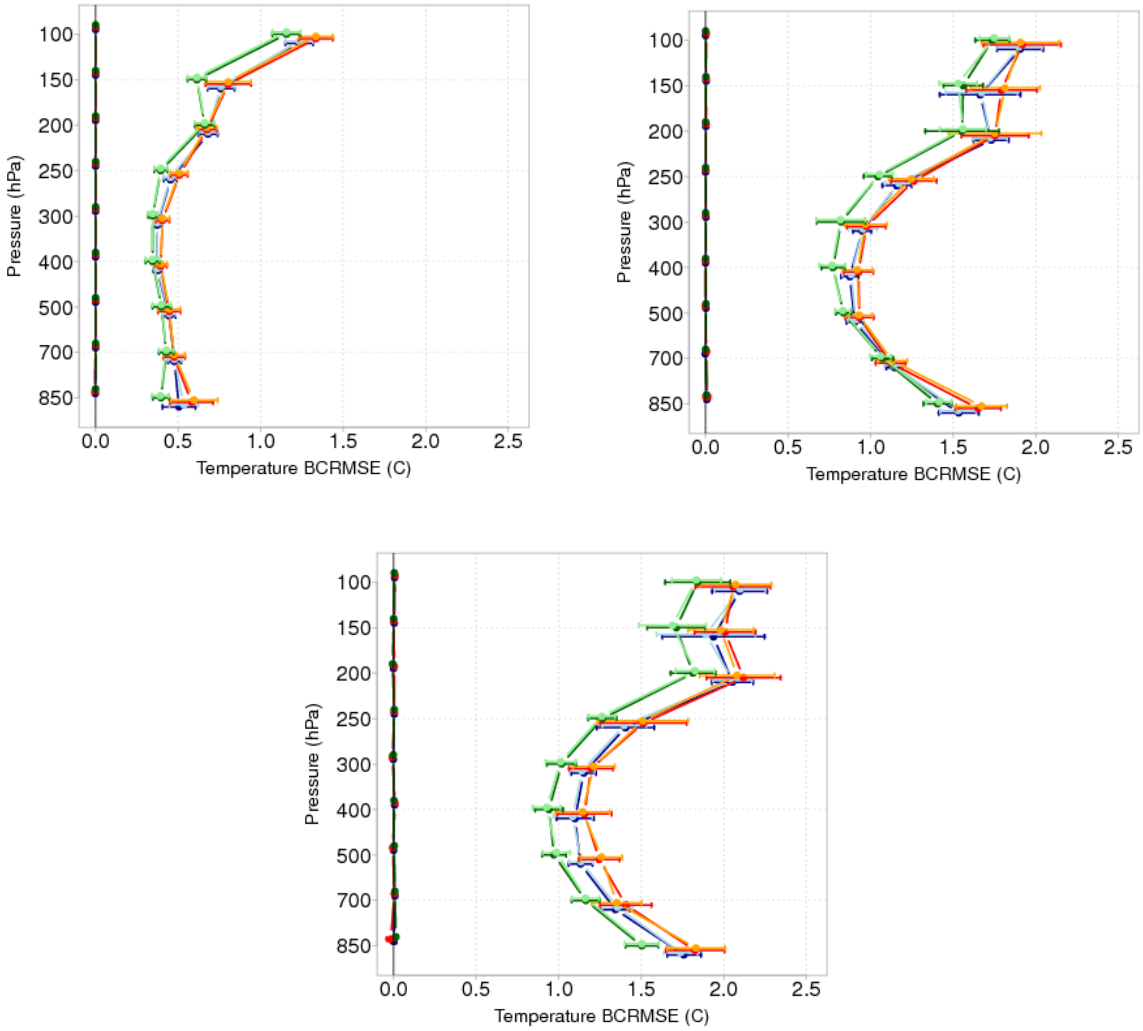

Figure 6. Vertical profile of the median BC-RMSE of temperature (C) at the 0-hour (top left), 12-h (top right), and 24-h lead times (bottom). The CONUS domain is shown in blue (ARW1=dark, ARW2=light), CONUS-East in green (ARW1=dark, ARW2=light) and CONUS-West in red for ARW1 and orange for ARW2. The ARW1-ARW2 pair-wise differences are also plotted in the corresponding colors for each domain and fall near the zero line. The horizontal bars represent the 99% CIs.

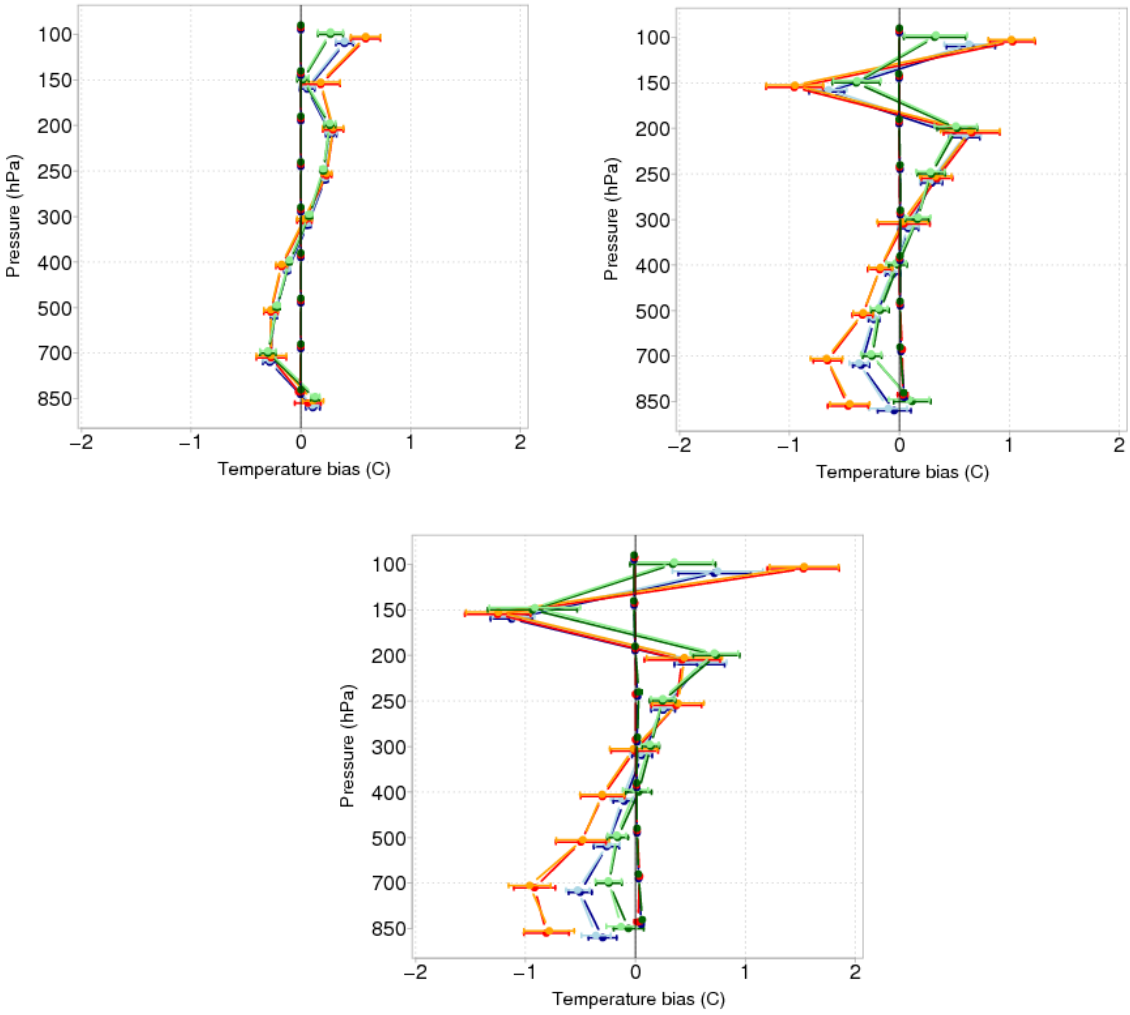

Figure 7. Same as Figure 6 except for bias.

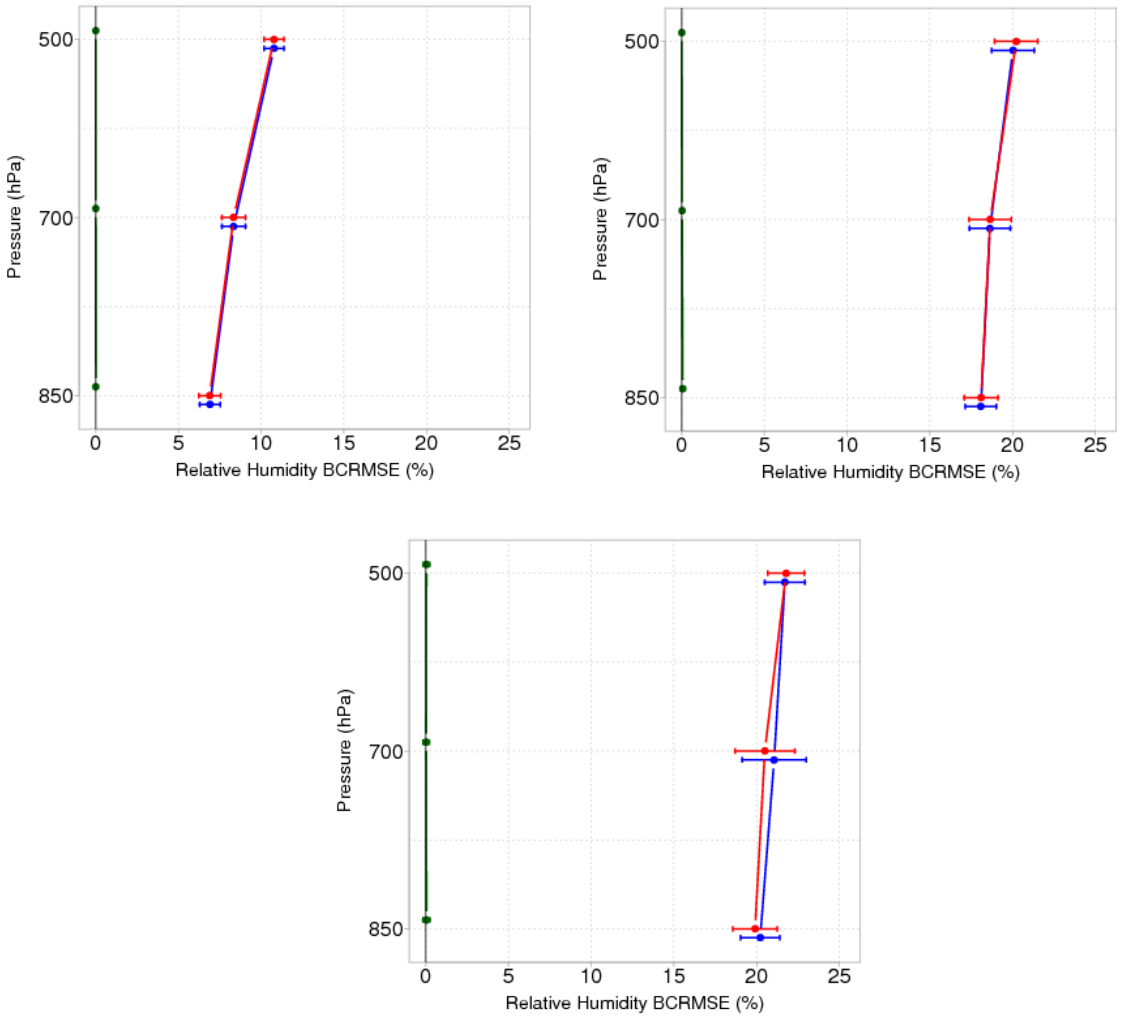

Figure 8. Vertical profile of the median BC-RMSE of Relative Humidity (%) for the CONUS domain only at the 0-hour (top left), 12-h (top right), and 24-h lead times (bottom). ARW1 is blue, ARW2 is red, and the ARW1-ARW2 pair-wise difference is green. The horizontal bars represent the 99% CIs.

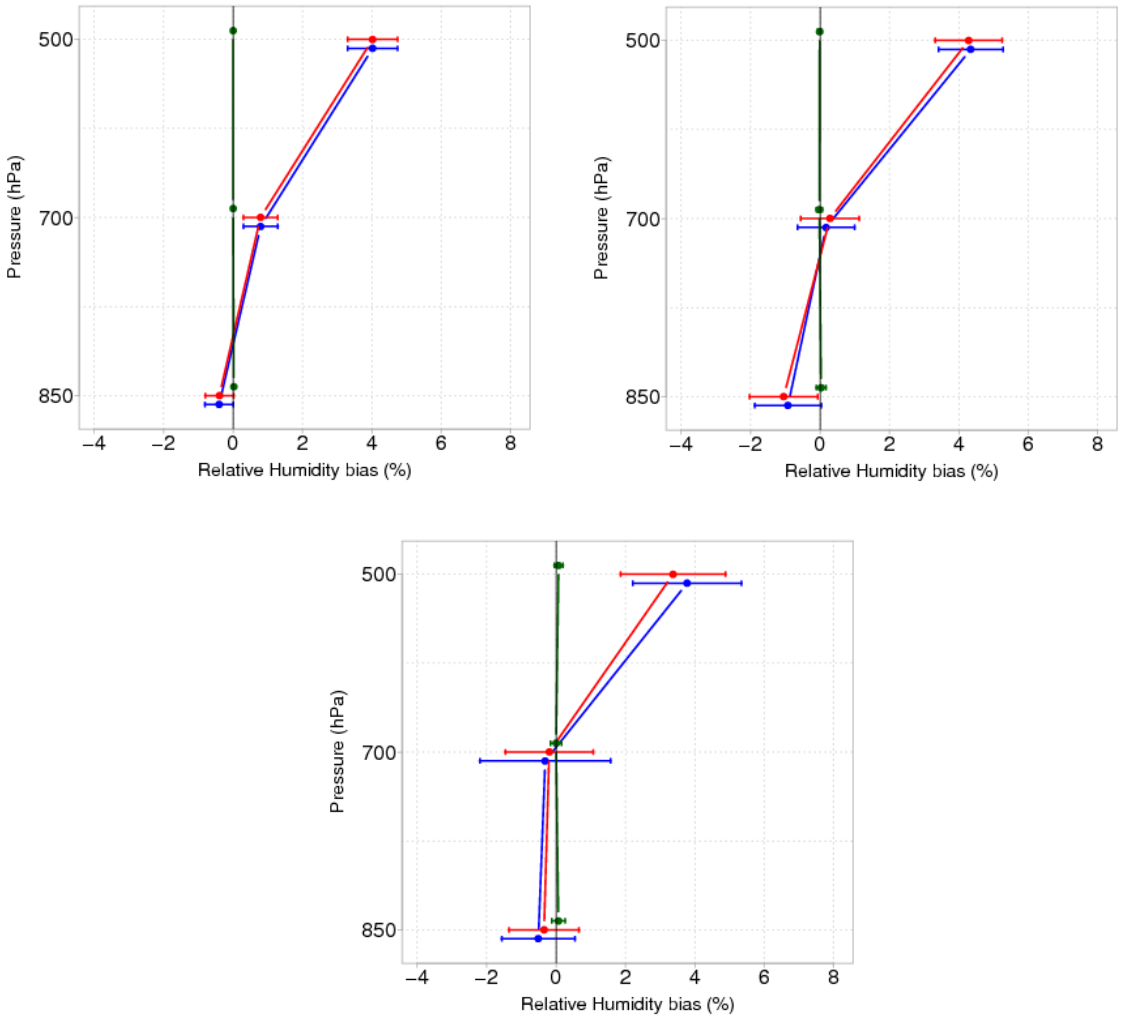

Figure 9. Same as Figure 8 except for bias.

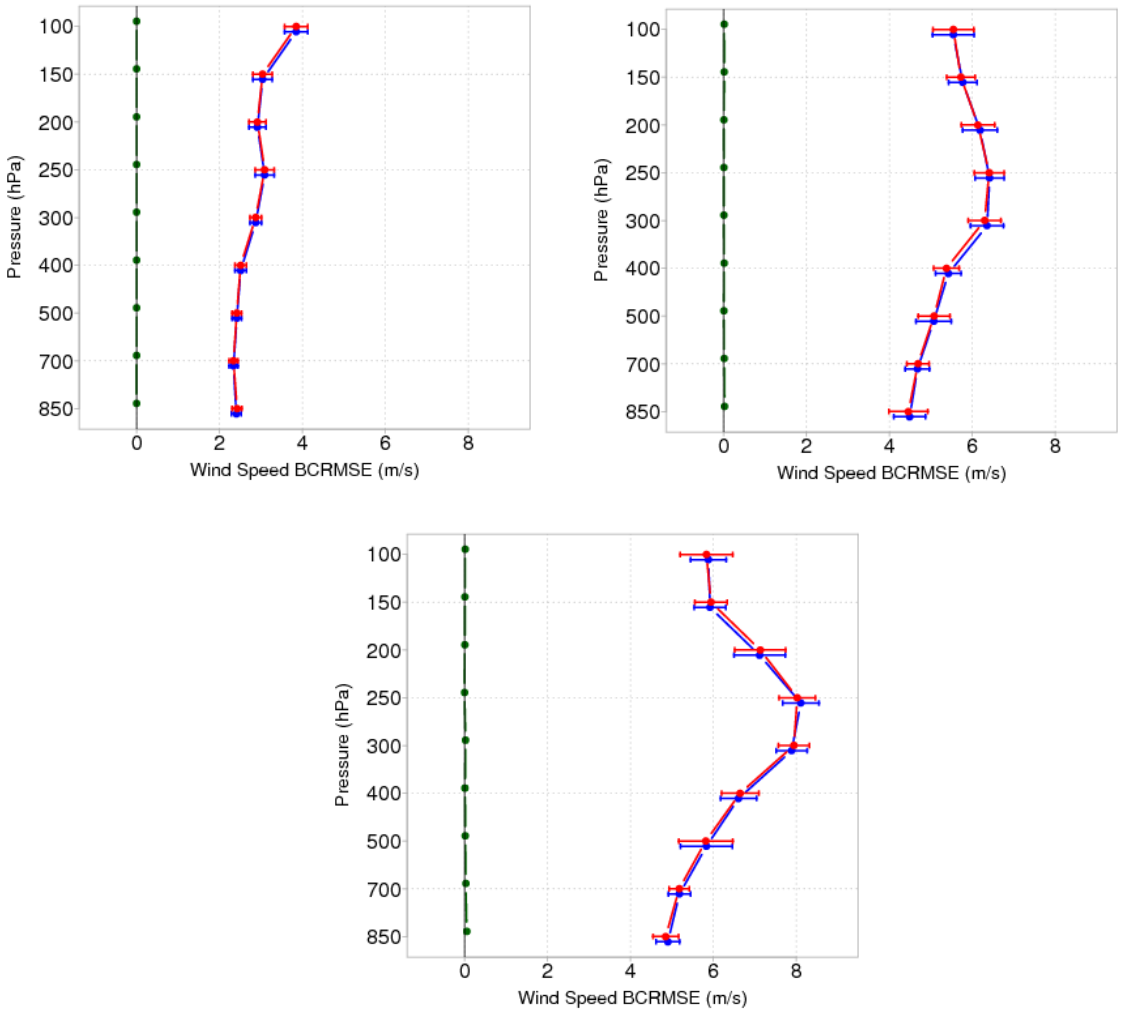

Figure 10. Vertical profile of the median BC-RMSE of Wind Speed (m/s) for the CONUS domain only at the 0-hour (top left), 12-h (top right), and 24-h lead times (bottom). ARW1 is blue, ARW2 is red, and the ARW1-ARW2 pair-wise difference is green. The horizontal bars represent the 99% CIs.

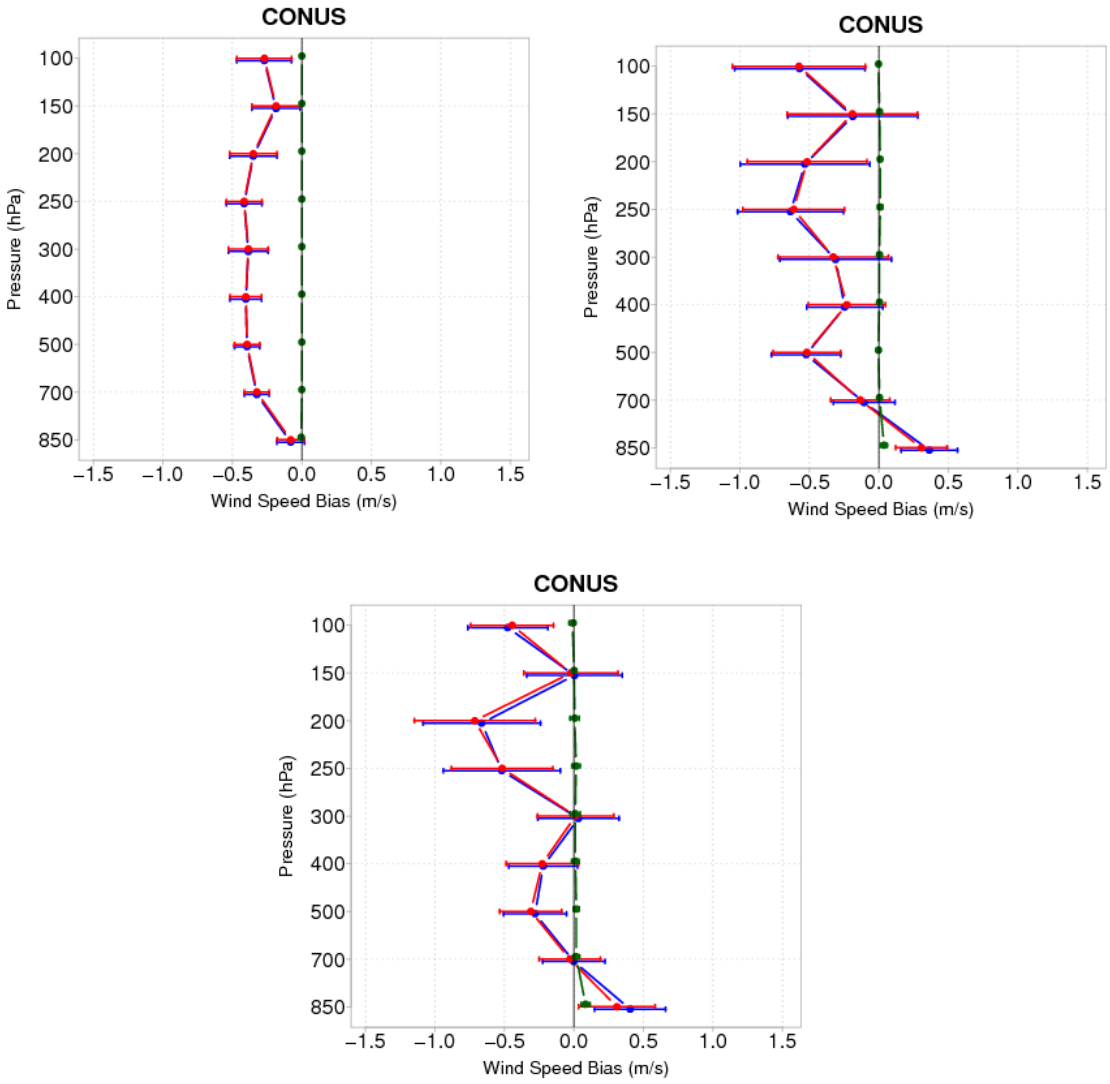

Figure 11. Same as Figure 10 except for bias.

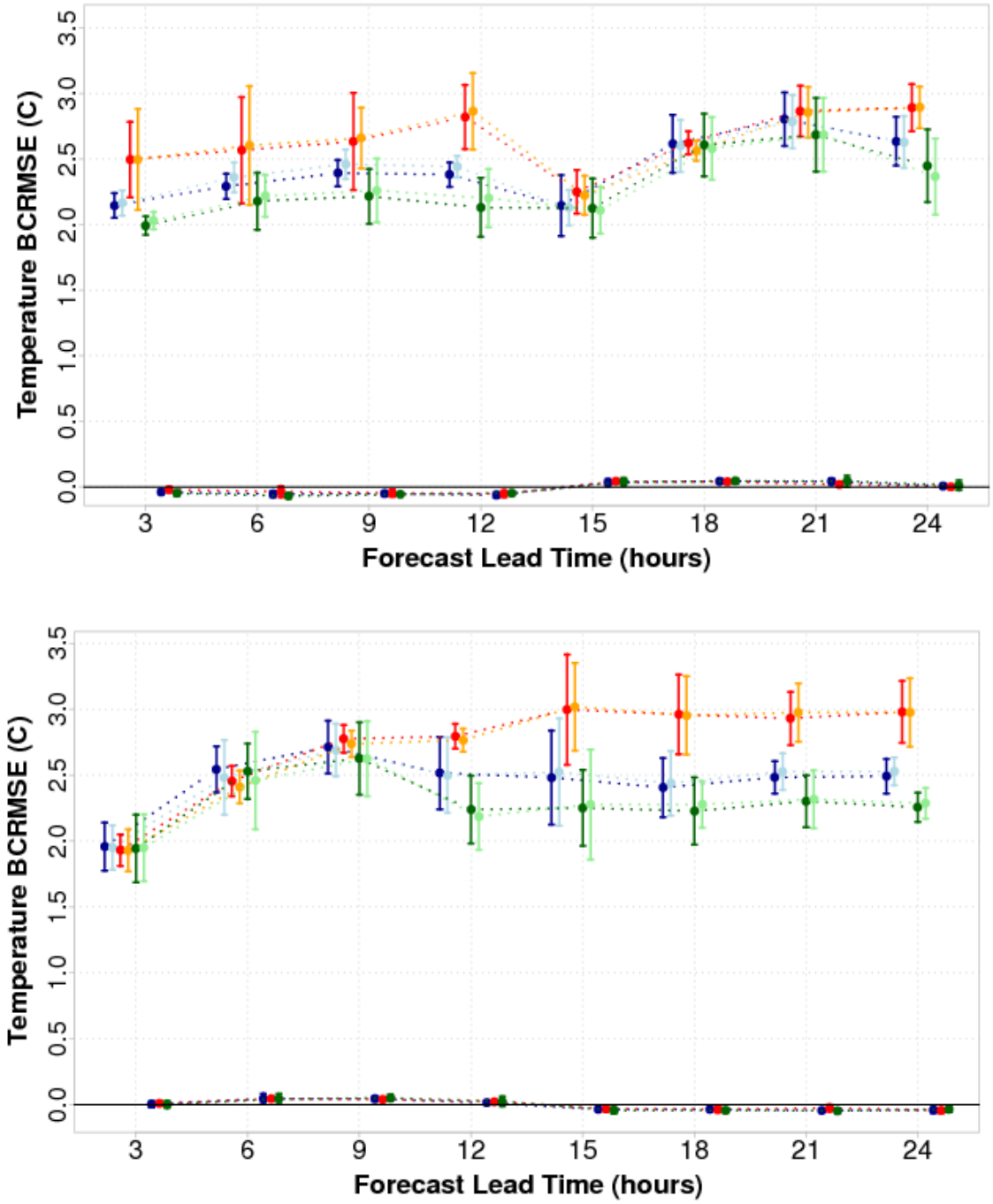

Figure 12. Time series plot of 2-m AGL temperature (C) for median BC-RMSE for the 00 UTC initializations only (top) and 12 UTC initializations only (bottom). The CONUS domain is shown in blue (ARW1=dark, ARW2=light), CONUS-East in green (ARW1=dark, ARW2=light) and CONUS-West in red for ARW1 and orange for ARW2. The pair-wise differences are also plotted in the corresponding colors for each domain and fall near the zero line. The vertical bars represent the 99% CIs.

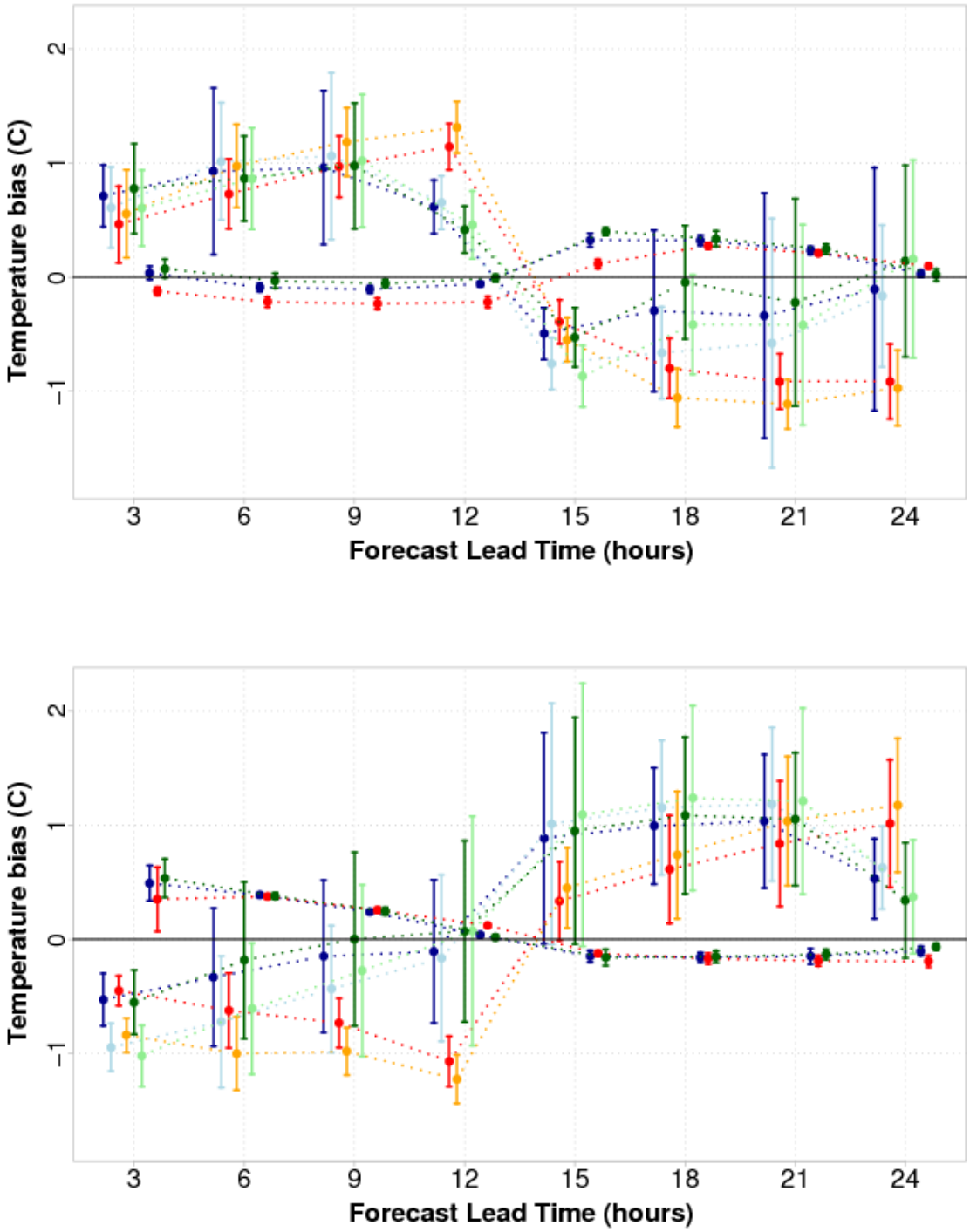

Figure 13. Same as Figure 12 except for bias.

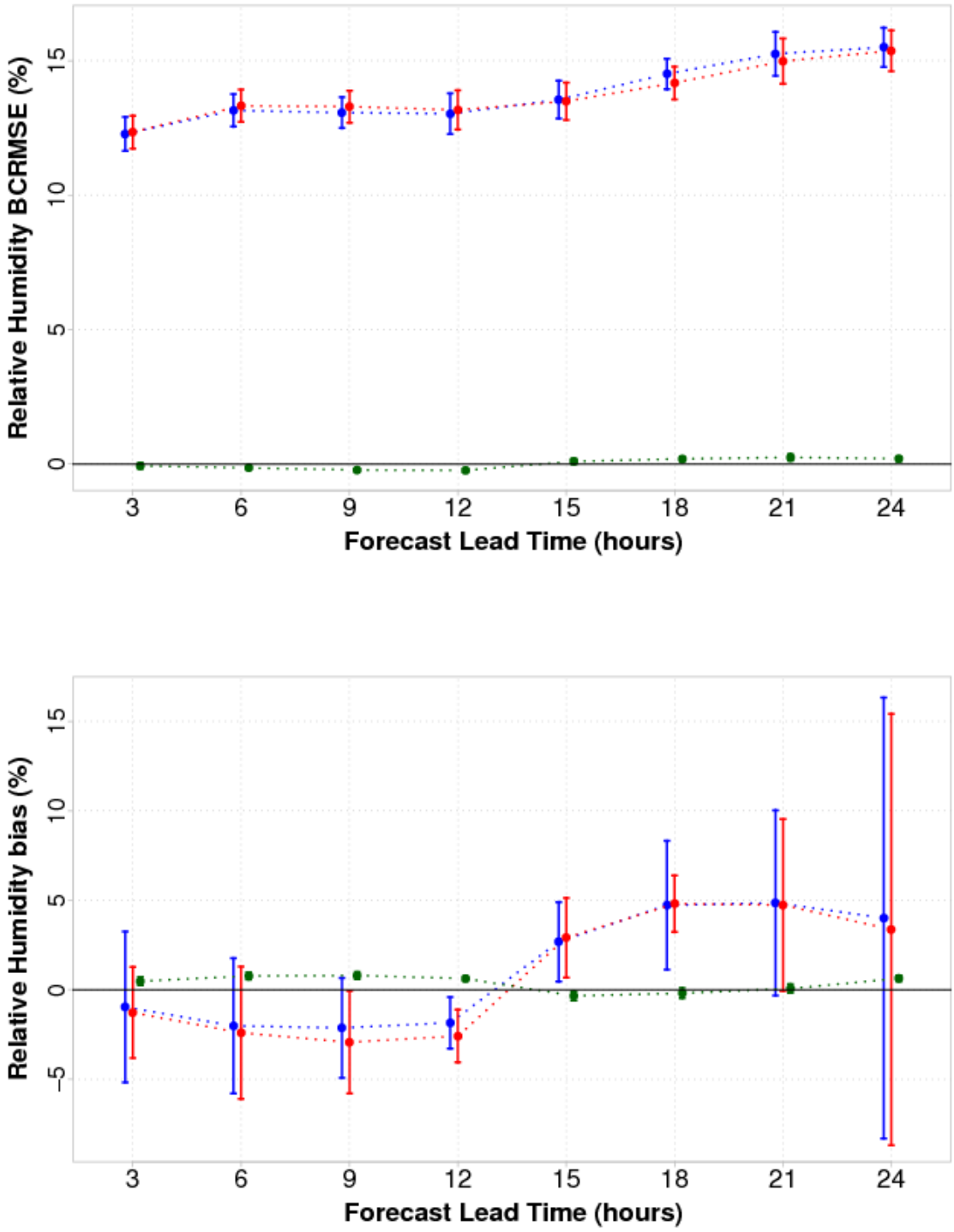

Figure 14. Time series plot of 2-m AGL relative humidity (%) for median BC-RMSE (top) and bias (bottom) for 00 UTC initializations and the CONUS domain only. ARW1 is blue, ARW2 is red, and the pair-wise difference is green. The vertical bars represent the 99% CIs.

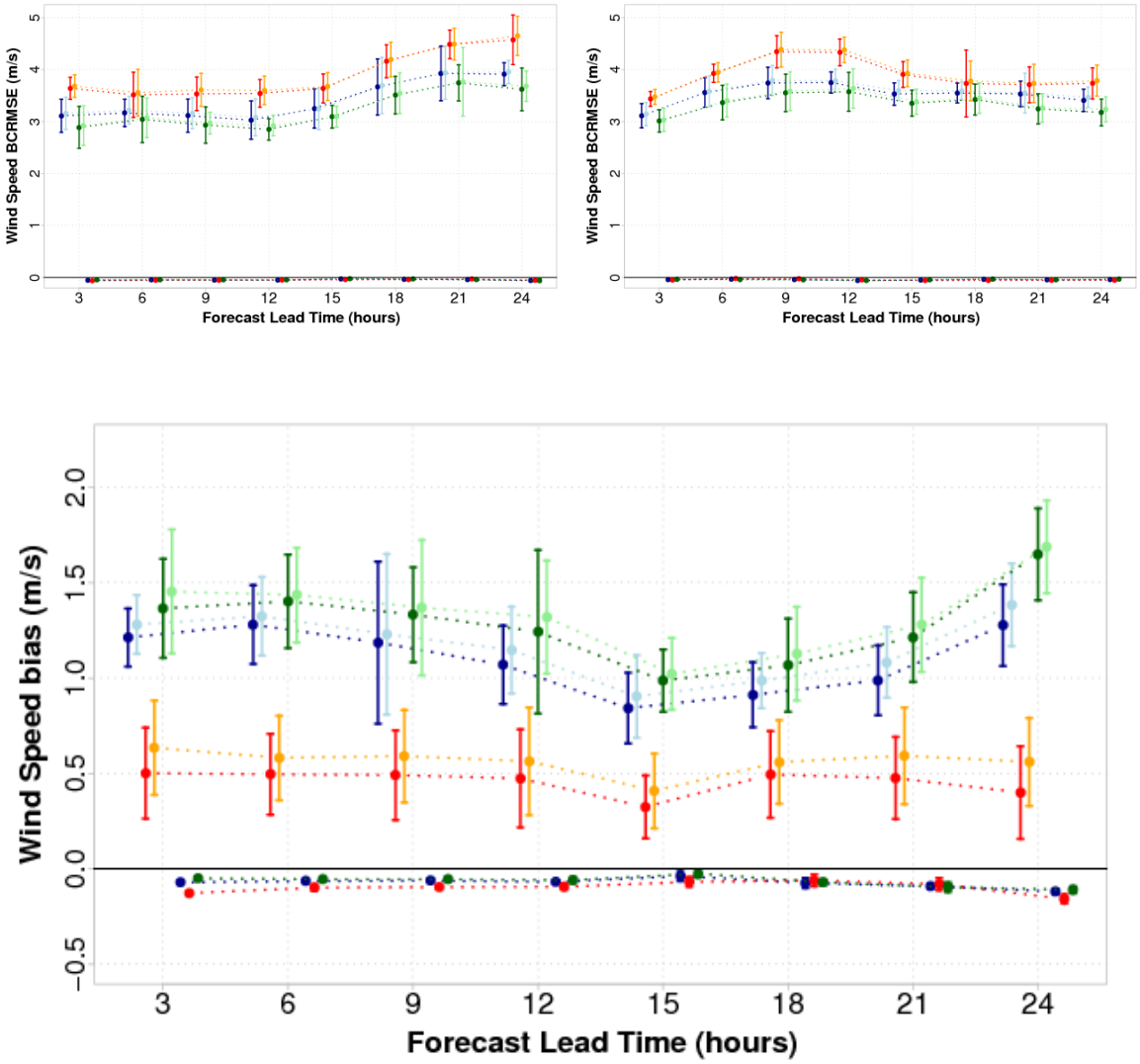

Figure 15. Time series plot of Median BC-RMSE for the 10-m AGL vector winds (m/s) for 00 UTC initializations (top left) and 12 UTC initializations (top right) and of median bias for the 10-m AGL wind speed (m/s) for the 00 UTC initializations only (bottom). The CONUS domain is shown in blue (ARW1=dark, ARW2=light), CONUS-East in green (ARW1=dark, ARW2=light) and CONUS-West in red for ARW1 and orange for ARW2. The pair-wise differences are also plotted in the corresponding colors for each domain and fall near the zero line. The vertical bars represent the 99% CIs.

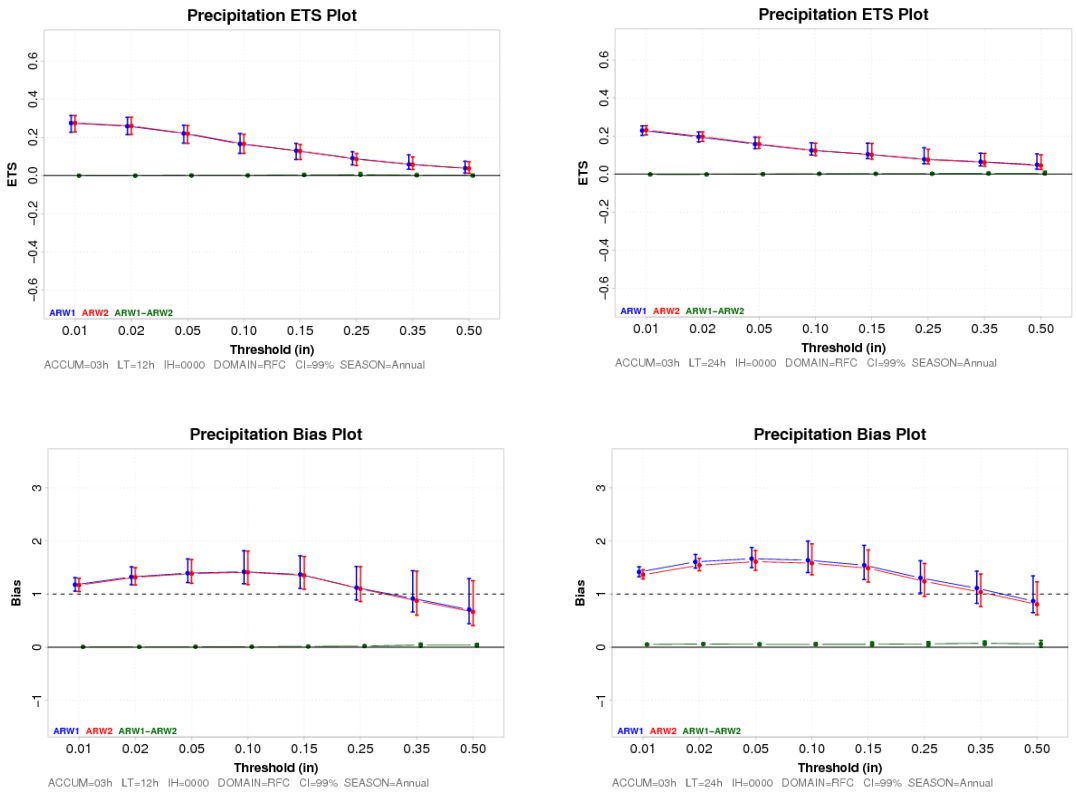

Figure 16. Equitable threat score (top row) and frequency bias (bottom row) for 3-hour accumulations at a lead time of 12 hours (left column) and 24 hours (right column) for 00 UTC initializations and the CONUS domain only. ARW1 is blue, ARW2 is red, and the pair-wise difference is green. The vertical bars represent the 99% CIs.

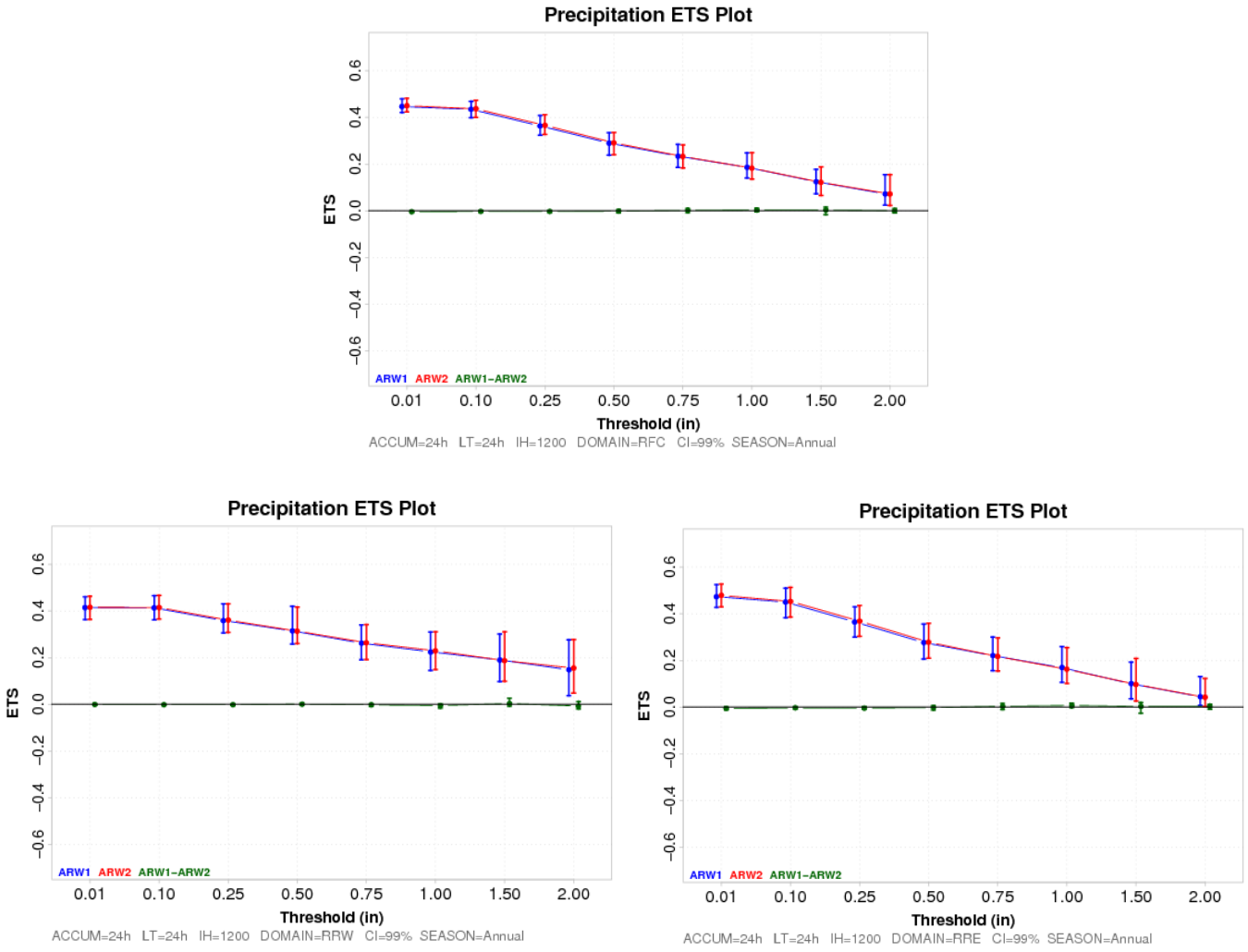

Figure 17. Equitable threat score for daily (24-hour) accumulations for the 12 UTC initializations only for the CONUS (top), CONUS-West (bottom left) and CONUS-East (bottom right) domains where ARW1 is blue, ARW2 is red, and the pair-wise difference is green. The vertical bars represent the 99% CIs.

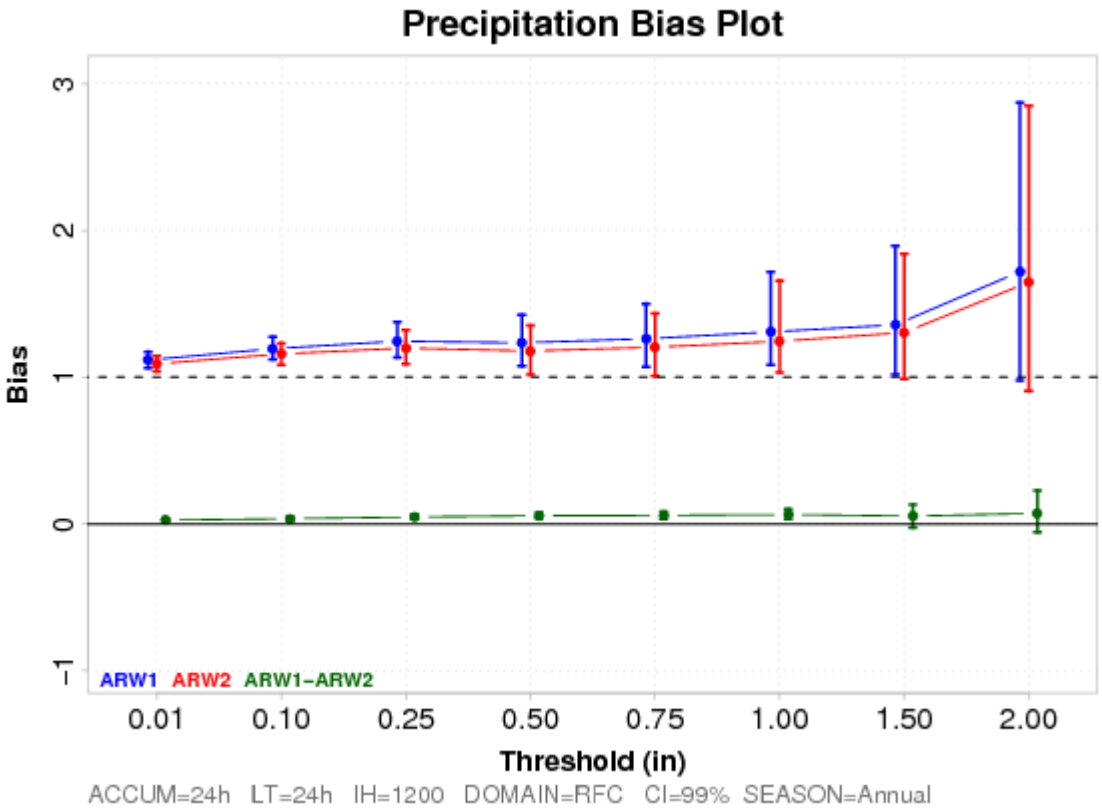

Figure 18. Frequency bias of the daily (24-hour) accumulations for the 12 UTC initializations and CONUS domain only where ARW1 is blue, ARW2 is red, and the pair-wise difference is green. The vertical bars represent the 99% CIs.

Appendix A: Subset of WRF ***namelist.input***

```
&time_control
 run_hours                  = 24,
 interval_seconds           = 10800
 history_interval           = 180,
 frames_per_outfile         = 1,
 io_form_history            = 2
/

 &domains
 time_step                  = 72,
 time_step_fract_num        = 0,
 time_step_fract_den        = 1,
 max_dom                    = 1,
 e_we                       = 400,
 e_sn                       = 304,
 e_vert                     = 51,
 num_metgrid_levels         = 40
 dx                         = 13300,
 dy                         = 13300,
 grid_id                    = 1,
 smooth_option              = 0
 p_top_requested            = 7500.
 ARW1:
          eta_levels   =   1.0000, 0.9980, 0.9940, 0.9870, 0.9750,
          0.9590, 0.9390, 0.9160, 0.8920, 0.8650, 0.8350, 0.8020,
          0.7660, 0.7270, 0.6850, 0.6400, 0.5920, 0.5420, 0.4970,
          0.4565, 0.4205, 0.3877, 0.3582, 0.3317, 0.3078, 0.2863,
          0.2670, 0.2496, 0.2329, 0.2188, 0.2047, 0.1906, 0.1765,
          0.1624, 0.1483, 0.1342, 0.1201, 0.1060, 0.0919, 0.0778,
          0.0657, 0.0568, 0.0486, 0.0409, 0.0337, 0.0271, 0.0209,
          0.0151, 0.0097, 0.0047, 0.0000,
 ARW2:
          eta_levels   =   1.0000, 0.9940, 0.9850, 0.9730, 0.9580,
          0.9430, 0.9270, 0.9100, 0.8900, 0.8650, 0.8350, 0.8020,
          0.7660, 0.7270, 0.6850, 0.6400, 0.5920, 0.5420, 0.4970,
          0.4565, 0.4205, 0.3877, 0.3582, 0.3317, 0.3078, 0.2863,
          0.2670, 0.2496, 0.2329, 0.2188, 0.2047, 0.1906, 0.1765,
          0.1624, 0.1483, 0.1342, 0.1201, 0.1060, 0.0919, 0.0778,
          0.0657, 0.0568, 0.0486, 0.0409, 0.0337, 0.0271, 0.0209,
          0.0151, 0.0097, 0.0047, 0.0000,

/

 &physics
 mp_physics                 = 8,
 ra_lw_physics              = 1,
 ra_sw_physics              = 1,
 radt                       = 30,
 sf_sfclay_physics          = 2,
 sf_surface_physics         = 3,
 bl_pbl_physics             = 2,
 bldt                       = 0,
```

```
 cu_physics                      = 3,
 cudt                           = 0,
 isfflx                         = 1,
 ifsnow                          = 1,
 icloud                          = 1,
 surface_input_source                = 1,
 num_soil_layers                   = 6,
 ucmcall                         = 0,
 mp_zero_out                       = 2,
 maxiens                         = 1,
 maxens                          = 3,
 maxens2                          = 3,
 maxens3                          = 16,
 ensdim                          = 144,
 slope_rad                        = 0,
 topo_shading                      = 0,
 /

 &dynamics
 w_damping                         = 1,
 diff_opt                       = 1,
 km_opt                          = 4,
 diff_6th_opt                     = 0,
 diff_6th_factor                   = 0.12,
 base_temp                        = 290.
 damp_opt                         = 1,
 zdamp                           = 5000.,
 dampcoef                         = 0.02,
 khdif                          = 0,
 kvdif                          = 0,
 non_hydrostatic                    = .true.,
 pd_moist                         = .true.,
 pd_scalar                        = .false.,
/

 &bdy_control
 spec_bdy_width                     = 5,
 spec_zone                        = 1,
 relax_zone                       = 4,
 specified                       = .true.,
 nested                          = .false.,
 /

 &namelist_quilt
 nio_tasks_per_group = 0,
 nio_groups = 1,
 /
```

Appendix B: Verification metrics

Root mean square error –

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2}$$

Answers the question: *What is the average magnitude of the forecast errors?*

Range: 0 to infinity. **Perfect score:** 0.

Characteristics: Simple, familiar. Measures "average" error, weighted according to the square of the error. Does not indicate the direction of the deviations. The *RMSE* puts greater influence on large errors than smaller errors, which may be a good thing if large errors are especially undesirable, but may also encourage conservative forecasting.

Bias-corrected root mean square error –

$$BCRMSE = \sqrt{s_{f-o}^2} = \sqrt{(s_f^2 + s_o^2 - 2s_f s_o r_{fo})}$$

Answers the question: *What is the standard deviation of the forecast errors?*

Range: 0 to infinity. **Perfect score:** 0.

Characteristics: Removes the effect of overall bias from the forecast-observation squared differences.

Bias – the correspondence between the mean forecast and mean observation (Mean Error) –

$$Mean\ Error = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)$$

Answers the question: *What is the average forecast error?*

Range: minus infinity to infinity. **Perfect score:** 0.

Characteristics: Simple, familiar. Also called the (additive) bias. Does not measure the magnitude of the errors. Does not measure the correspondence between forecasts and observations, i.e., it is possible to get a perfect score for a bad forecast if there are compensating errors.

Bias score (frequency bias) –

$$BIAS = \frac{hits + false\ alarms}{hits + misses}$$

Answers the question: *How did the forecast frequency of "yes" events compare to the observed frequency of "yes" events?*

Range: 0 to infinity. **Perfect score:** 1.

Characteristics: Measures the ratio of the frequency of forecast events to the frequency of observed events. Indicates whether the forecast system has a tendency to underforecast (*BIAS*<1) or overforecast (*BIAS*>1) events. Does not measure how well the forecast corresponds to the observations, only measures relative frequencies.

Equitable threat score (Gilbert skill score) –

$$ETS = \frac{hits - hits_{random}}{hits + misses + false\ alarms - hits_{random}}$$

where $$hits_{random} = \frac{(hits + misses)(hits + false\ alarms)}{total}$$

Answers the question: *How well did the forecast "yes" events correspond to the observed "yes" events (accounting for hits due to chance)?*

Range: -1/3 to 1, 0 indicates no skill. **Perfect score:** 1.

Characteristics: Measures the fraction of observed and/or forecast events that were correctly predicted, adjusted for hits associated with random chance (for example, it is easier to correctly forecast rain occurrence in a wet climate than in a dry climate). The *ETS* is often used in the verification of rainfall in NWP models because its "equitability" allows scores to be compared more fairly across different regimes. Sensitive to hits. Because it penalizes both misses and false alarms in the same way, it does not distinguish the source of forecast error.