Weather Research and Forecasting Core Test

DTC Report

Louisa Bogar Nance

Contributors: The WRF Core Test was a large undertaking that involved contributions from multiple organizations. The following individuals contributed to the production of the retrospective runs and/or the DTC's evaluation of these runs:

DTC - Ligia Bernardet, Meral Demirtas, Robert Gall, Steve Koch

NOAA/ESRL – Stan Benjamin, John Brown, Randy Collander, Chris Harrop, Andrew Loughe, Tanya Smirnova

NCAR/RAL - Tressa Fowler, Greg Thompson

University of Colorado/CIRES – Greg Noonan, Betsy Weatherhead

Executive Summary

The current Weather and Research Forecasting (WRF) Software Framework (WSF) supports two dynamical solvers: the Advanced Research WRF (ARW) and the Nonhydrostatic Mesoscale Model. The goal of the WRF Core Test was to determine the impact of these dynamical solvers on the forecast. With this intent, parallel runs of WRF using the two dynamical cores were configured in a way to limit the configuration differences to those related to the dynamical solvers. The retrospective runs produced for the WRF Core Test underwent both objective and subjective evaluation. The DTC Core Test report focuses on the differences between the standard verification measures for the two dynamical cores. Only the differences for which an assessment of statistical significance has been completed are discussed in this report. The following points summarize the differences in the verification statistics for the two dynamical cores that are statistically significant and show consistency across season, physics package, and observational data type:

- The ARW vector wind RMSE is less than the NMM vector wind RMSE in the 300-150 hPa layer. The magnitude of these differences also meets the criteria for concern.
- Differences between the wind speed bias for the two dynamical cores indicate the upper level wind in the ARW forecasts are generally weaker than those in the NMM forecasts, whereas the lower level winds in the ARW forecasts tend to be stronger than those in the NMM forecasts.
- The ARW temperature RMSE is less than the NMM temperature RMSE in the 400-200 hPa layer. Although the differences in this layer show consistency across physics packages and observational data type, the magnitude of these differences decreases during the second 12 hours of the forecast. The magnitude of the differences in this layer generally does not exceed the threshold for concern.
- Differences between the temperature bias for the two dynamical cores indicate the ARW forecasts are generally colder than the NMM forecasts. Both cores exhibit a negative temperature bias at lower levels that transitions to a positive bias at upper levels. This vertical structure combined with the colder temperatures in the ARW forecasts leads to the magnitude of the temperature bias being smaller for the NMM at lower levels and smaller for the ARW at upper levels.

- The relative humidity RMSE differences did not exhibit any statistically significant signature that was consistent for both physics packages. On the other hand, a number of the differences for phase 2 and one difference for phase 1 that were found to be statistically significant did exceed the thresholds for concern and serious concern. The ARW relative humidity RMSE was smaller than that of the NMM for all of these cases.
- Differences between the relative humidity bias for the two dynamical cores indicate the ARW forecasts are generally associated with lower values of relative humidity than the NMM forecasts. This tendency for lower relative humidity in the ARW forecasts can at least partially be explained by the tendency for the temperatures to be colder in the ARW forecasts.
- Very few of the differences between the QPF verification measures for the two dynamical cores are statistically significant and show consistency for the two physics packages. Only the bias differences at the lowest thresholds produce consistent statistically significant results. The bias differences at the lowest thresholds indicate the NMM produces less overestimation of the areal coverage than the ARW. On the other hand, all of these differences do not exceed the threshold for concern.

1. Introduction

The current Weather and Research Forecasting (WRF) Software Framework (WSF) supports two dynamical solvers: the Advanced Research WRF (ARW, Skamarock et al. 2005) developed by the Mesoscale and Microscale Meteorology (MMM) Division of the National Center for Atmospheric Research (NCAR), and the Nonhydrostatic Mesoscale Model (NMM – Janjic 2003a,b) developed by the National Centers for Environmental Prediction (NCEP). WRF also offers a variety of physics packages. Parallel runs of these two dynamic solvers are available from a variety of experiments (e.g., WRF Test Plan, Spring Program, DTC Winter Forecast Experiment, etc…Seaman et al. 2004, Kain et al. 2005, Bernardet et al. 2005), but the differences between these parallel runs can not be solely attributed to differences between the two dynamical cores because the dynamical solvers were not configured to use the same physical parameterizations and/or initial conditions. The WRF Developmental Testbed Center (DTC) and the Global Systems Division (GSD) of NOAA's Earth System Research Laboratory (ESRL) addressed the need for a controlled comparison of these two dynamical cores through intensive retrospective testing (i.e., parallel runs of the two dynamical solvers using initial and lateral boundary conditions based on the same input data, as well as the same suite of physics parameterizations). The goal of the WRF Core Test was to determine the impact of the dynamical solvers on the forecast. With this intent, parallel runs of WRF using the two dynamical solvers were configured such differences between the forecasts would be confined to those related to the dynamical solvers, at least within the limitations of the current end-to-end system. This intensive testing provides much needed information for the upcoming dynamical core recommendation to be made by ESRL-GSD for the WRF Rapid Refresh to be run at NCEP.

The parallel runs produced for the WRF Core Test underwent three basic types of evaluation: objective verification based on the standard verification measures, subjective case studies investigating systematic differences, and aviation-specific evaluation (both case studies and statistics) by the Product Development Teams (PDTs – Convective, Icing, Ceiling and Visibility, and Turbulence) of the Aviation Weather Research Program (AWRP). This report focuses on the set-up of the experiment and the standard verification measures.

2. Experiment Design

2.1 Forecast Cycles

The WRF Core Test focused on four month-long retrospective time periods corresponding to the four seasons:

Summer:	15 July – 15 August 2005
Autumn:	1 – 30 November 2005
Winter:	15 January – 15 February 2006
Spring:	25 March – 25 April 2006

Twice daily (00 and 12 UTC cycles) 24-h forecasts for each retrospective period were generated on a 13-km grid for each dynamical core with output files every three hours. For the 12 UTC cycle on 21 July 2005, the WSF auxillary output option was utilized to produce hourly output for a select subset of two-dimensional fields to meet the needs of the Convective AWRP PDT.

Forecast cycles for which the end-to-end system was not able to run to completion are summarized in Appendix A. Only 22 of the potential 252 forecast cycles were not able to run completely from initialization to verification. Fourteen of these incomplete cycles were related to missing input data. Only four of the incomplete cycles were the result of the model failing to produce a 24-h forecast. These forecast failures are still under investigation. The four remaining cycles encountered problems in the post-processing stage due to small soil moisture values. The verification results discussed in section 3 are based on only those forecast cycles for which the end-to-end system ran to completion for both dynamical cores and both physics package configurations.

2.2 Input data

The same gridded data were used to generate initial and lateral boundary conditions for each dynamical core. RUC13 grids were used to produce the initial conditions (ICs), whereas NAM212 grids were used to produce the lateral boundary conditions (LBCs). This mixture of input data was chosen because RUC13 grids were not available out to 24 hours for all of the retrospective time periods. To maintain consistency between the RUC13 ICs and NAM LBCs, the NAM grids from the prior 06 UTC and 18 UTC cycle were used to produce LBCs for the 12 UTC and 00 UTC forecast cycles, respectively. In addition, the sea surface temperature (SST) field for the forecasts was obtained from NCEP's daily, real-time, SST product, which 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 used for this testing did not include a data assimilation component, the RUC13 cloud fields were included in the initial conditions, so the runs considered in this study are not truly cold start forecasts.

2.3 Domains

The ARW offers three map-projection options: Lambert-conformal, polar stereographic or Mercator, whereas the NMM uses a rotated latitude-longitude map projection. Hence, the two dynamical cores can not be run on an identical grid. The Lambert-Conformal map projection was used for the ARW runs considered in this study. The domain for each core was selected such that it fit within the RUC13 domain, while attempting to minimize the following differences between domain characteristics for each core: the number of grid points, the range of grid spacing across the domain, and the locations of the lateral boundaries (see Fig. 1). For the selected domains, the number of computational grid points differs by 0.014% (i.e., the ARW domain has 0.014% more computational grid points than the NMM domain). The range of grid spacing for the ARW is 12.68 to 13.47 km, whereas the range for the NMM is 13.07 to 13.41 km.

The two dynamical cores also use different vertical coordinates. Hence, the two cores could not be setup to run on identical vertical levels. Both cores were configured to use 50 vertical levels with similar, but not identical, vertical spacing.

2.4 Pre-Processing

Each dynamical core has its own pre-processing software package, referred to as the Standard Initialization or SI. The approaches to processing the input data are not necessarily the same in these two packages. Hence, differences between the initial and lateral boundary conditions, as

well as the static fields (e.g. terrain height) that go beyond those stemming from differences in the grid projection/staggering for the two dynamical cores are possible. The differences in the initial conditions are assessed by considering the differences between the verification statistics of the two dynamical cores at the initial time.

The version of the NMM SI used for the WRF Core Test applies a 5-point smoother to the 30-second terrain data during the domain localization process, whereas the ARW SI utilizes a method where the terrain is defined on four different successive grids and the degree of smoothing is controlled by two namelist parameters: SILAVWT_PARM_WRF and TOPTWVL_PARM_WRF. The ARW terrain was created using SILAVWT_PARM_WRF=0 and TOPTWVL_PARM_WRF=4 to generate a reasonable match between the terrain forcing for the two dynamical cores. Figure 2 shows the difference between the terrain fields that were used in the retrospective forecasts.

2.5 Model

The inherent differences between the numerics for the two dynamical solvers dictate using different time-steps for the same grid spacing. The time-steps selected for the WRF Core Test are based on recommendations from the respective developers, as well as real-time runs of each dynamical core leading up to the start of the retrospective runs. The NMM used a time-step of 30 s, whereas, the ARW used a large time-step of 72 s and a small time-step of 18 s. In the interest of maintaining consistency between the two dynamical core configurations, the physics routines were called at the same frequency when possible, and close to the same frequency when an exact match was not possible. Both cores called the radiation package every 30 minutes, whereas the ARW called the remaining physics packages every large time step (72 s) and the NMM called these packages every other time step (60 s). An exact match for the non-radiation physics packages was not possible because the ARW solver currently does not offer an option to call the microphysics package less frequently than every time step.

The ARW solver also offers a number of run-time options for the numerics, as well as various filter and damping options (Skamarock et al 2005). 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 ARW was configured to use the following filter/damping options: 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, and a 5-km-deep diffusive damping layer at the top of the domain (coefficient 0.02).

Each dynamical core was run with the same physics package configuration. Two physics package configurations were considered:

	Phase 1	Phase 2
Microphysics:	Ferrier	Thompson
Planetary Boundary Layer:	Mellor-Yamada-Janjic	Mellor-Yamada-Janjic
Convection:	Betts-Miller-Janjic	Grell-Devenyi
Land-Surface Model:	Noah (option 99)	RUC

Beyond the physics packages, the same basic configuration was used for both phase 1 and phase 2 of the testing, except that the phase 2 runs utilized the microphysics option mp_zero_out with the threshold set to 1e-12. The version of this option used in the phase 2 runs sets negative water vapor values to zero and all moist array values that fall below the specified threshold to zero.

2.6 Post-Processing

The WRF-POST (Chuang et al. 2004), developed at NCEP, was used to post-process the retrospective forecasts. The output for each dynamical core was interpolated from their respective native horizontal grid to the non-staggered RUC13 grid. Points of the RUC13 grid that fell outside the WRF grids were filled with missing values. Three types of post-processed files were produced for this experiment: 1) three-dimensional fields on constant pressure levels and two-dimensional fields, 2) three-dimensional fields on the model native vertical levels, and 3) two-dimensional fields. The files containing the pressure-level and two-dimensional fields were used to compute verification statistics and produce graphics, whereas the native level output was made available to the AWRP PDTs for evaluation using their own diagnostic tools.

2.6 Verification

2.6.1 Surface and upper air

The retrospective forecasts were evaluated through objective verification of surface and upper air fields using the NCEP Verification System (Chuang et al. 2004). The measures considered in this report are root-mean-square-error (RMSE) and bias. The NCEP system employs a grid-to-point verification approach in which forecast fields are bi-linearly interpolated to the observation location. The system produces upper air verification statistics for radiosonde observations (temperature, relative humidity, and winds) at mandatory levels and aircraft observations (temperature and winds) for a number of layers. Given the known errors associated with radiosonde humidity measurements at higher altitudes, our analysis of the upper air relative humidity verification statistics focuses on the levels between 850 and 500 hPa. At the surface, forecasts of 2-m temperature and relative humidity and 10-m winds are compared to METAR observations. The surface and upper air observations for this verification were obtained from RUC prepbufr files. Verification statistics were computed for three different domains: CONUS, West, and East (see Fig. 3). The CONUS verification domain is a Lambert-Conformal projection using the same parameters as the ARW computational domain with its outer limits defined to be 12 grid points in from the ARW domain boundary. The West and East domains are sub-domains chosen to isolate the western mountainous region from flatter region in the eastern United States.

An assessment of the statistical significance of the differences between the two dynamical core upper air verification statistics was carried out by Betsy Weatherhead and Greg Noonan of the Cooperative Institute for Research in the Environmental Science (CIRES) at the University of Colorado. Their assessment of statistical significance utilizes a pairwise comparison that applies an auto-correlation correction (for more information on this analysis, see Weatherhead et al 2006). Significance testing was done for all three verification domains for the annual average and for each seasonal average. For this report, the differences between the RMSE and bias for the two dynamical cores was considered statistically significant if the mean $\pm 4\sigma$, where σ represents the standard error, did not include zero. This criterion corresponds to a 99.9% confidence interval.

An assessment of the statistical significance of the differences between the two dynamical core surface verification statistics is not available at this time. Hence, a discussion of the surface verification statistics is not included in this report. A discussion of the surface verification results will be prepared as part of a separate report as soon as this statistical significance assessment is complete.

2.6.2 Precipitation

The precipitation forecasts were evaluated through objective verification using the NCEP Quantitative Precipitation Forecast (QPF) Verification System. The measures considered in this report are Equitable Threat Score (ETS) and bias. The NCEP system employs a grid-to-grid verification approach where the model forecast and observations are remapped to the same verification grid (Lin et al. 1999). The Eta218 grid was used for this verification grid. The package can compute verification statistics for 24-h and 3-h accumulations. The NCEP/CPC 1/8 degree daily precipitation analysis (accumulated from 12 UTC to 12 UTC) was used for verification of the 24-h accumulation, and the Stage-II national multi-sensor hourly precipitation analysis was used for verification of the 3-h accumulation. Given the retrospective forecasts were only integrated out to 24 hours, verification for the 24-h accumulation was only possible for the 12 UTC forecast cycles.

An assessment of the statistical significance of the differences between the two dynamical core QPF verification statistics for the CONUS 24-h accumulation was carried out by Tressa Fowler of NCAR's Research Applications Laboratory (RAL). An estimate of the uncertainity in the differences between the precipitation statistics for the two dynamical cores was obtained by applying a computer resampling method (Efron and Tibshirani 1994). This method samples from the daily contingency tables of each core for each precipitation threshold with replacement. Because of the paired nature of the core comparison, the contingency tables for a single time are selected for both cores. By accumulating the counts in the daily contingency tables over the entire period of interest, a single new contingency table for each core/threshold is derived. The statistics of interest (bias and ETS) are calculated from the contingency tables. The difference between these statistics for the two dynamical cores is accumulated. This process is repeated a large number of times (5000 in this case), yielding an empirical distribution of the difference in the statistics, from which the uncertainty in this difference can be estimated by finding the confidence bounds that separate 2.5% of the 5000 empirical samples into each tail.

Although QPF verification statistics are available for the seasonal averages, the verification sub-domains, and the 3-h accumulations, this report only discusses the results for the CONUS 24-h accumulation because an assessment of the statistical significance of the differences between the two dynamical cores for the other QPF verification statistics is not available at this time.

2.6.3 Threshold criteria

Prior to generating the retrospective runs and their corresponding error statistics, a guideline for error difference thresholds was proposed by Stan Benjamin. These proposed thresholds were based on his group's experience with RUC implementations and observation impact experiments. The proposed thresholds, which are summarized in Table 1, are generally consistent with the formal statistical-significance thresholds in Benjamin et al. (2004). Differences that fall in the green category are considered insignificant. Differences that fall in the yellow category are considered to be of concern. And finally, differences that fall in the red category are considered to be of serious concern.

3. Verification Results

3.1 Upper Air

RMSE is a positive quantity. Hence, the dynamical core with the smaller RMSE can be determined by simply subtracting one RMSE from the other and the sign of the difference indicates which dynamical core has the smaller RMSE. On the other hand, bias can be positive or negative, which makes interpreting differences less straight forward. Given forecast skill is the focus of this report (i.e., determining which dynamical core produces the bias of the smallest magnitude), the bias comparisons focus on which bias is smaller in magnitude. To facilitate this comparison, the absolute value of each bias was used to compute the differences between the two dynamical cores (i.e., |ARW bias| - |NMM bias|). All RMSE and bias differences were computed by subtracting the NMM verification measure from its ARW counterpart. Hence, positive differences are associated with the NMM verification measure being smaller than that for the ARW, and negative differences are associated with the ARW verification measure being smaller than that for the NMM.

3.1.1 Winds

3.1.1.1 RMSE

Forecast Hour 00

The initial vector wind RMSE for both dynamical cores ranges from approximately 2 to 4 ms^{-1} for the radiosonde data and 3 to 5 ms^{-1} for the aircraft data, with the largest RMSE occurring at upper levels (see Fig. 4a-b). The initial conditions for the summer season generally have the smallest RMSE, whereas the initial conditions for the winter season generally have the largest RMSE. Given the flow tends to be stronger in the winter season and the level of maximum RMSE for all seasons corresponds to the jet stream, this initial RMSE appears to scale with the strength of the flow. The RMSE differences for both data sets exhibit the same basic overall trend with height (i.e., the ARW RMSE is smaller than that for the NMM at lower levels, transitioning to the ARW RMSE being larger than that for the NMM at upper levels), but the magnitude of the differences for the aircraft data is smaller at upper levels (see Fig. 4c-d). The distinction can also be seen in the assessment of which differences are statistically significant. Most of the aircraft differences at upper are statistically significant, whereas only a few of the aircraft differences are statistically significant at these levels (see Fig. 4e-f). All of the initial RMSE differences fall in the green category, except for the radiosonde differences above 250 hPa, which meet the yellow criteria for all seasons except summer.

Forecast Hour 12

During the first 12 hours of the forecast, the vector wind RMSE increases at all levels, with the largest increase occurring at upper levels (see Fig. 5a-b). Summer once again has the smallest overall RMSE and winter the largest. The ARW RMSE is smaller than the NMM RMSE above 300 hPa, which is opposite that found at the initial time (see Fig. 5c-d). The RMSE differences for the radiosonde data at 200 hPa and the aircraft data in the 250-200 hPa layer are statistically significant for both phase 1 and 2, but only for the annual average, and the winter and spring (aircraft only) seasons (see Fig. 5e-f). Most of these statistically significant differences also meet the yellow criteria. The statistical significance of these differences also carries over to the verification sub-domains, except for the radiosonde winter season, which is only statistically significant for the East sub-domain (not shown).

The NMM RMSE is smaller than the ARW RMSE in the layer from 550 to 400 hPa (see Fig. 5c-d). The RMSE differences for the radiosonde data at 400 and 500 hPa are statistically significant for the annual average (phase 1 and 2), and summer (phase 1-500mb and 2) and spring (phase 2-500 hPa) seasons, whereas the differences for the aircraft data in the 550-400 hPa layer are only statistically significant for the annual average (see Fig. 5e-f). These RMSE differences all meet the green criteria. The significance of these differences only carries over to the East sub-domain (not shown).

The RMSE differences based on the radiosonde data suggest that the NMM RMSE also tends to be smaller than the ARW RMSE at 700 hPa (Fig. 5c), whereas the aircraft data suggest that the dynamical core with the smaller RMSE at this level depends on the season (Fig. 5d). Only the differences for the radiosonde data show any statistical significance at this level and the significance of these differences show less continuity across phase, season, and sub-domain than the differences for the other two layers (Fig. 5e-f).

Forecast Hour 24

During the second 12 hours of the forecast, the vector wind RMSE continues to increase at all levels, with the largest increase once again occurring at upper levels (see Fig. 6a-b). The ARW RMSE continues to be smaller than the NMM RMSE above 300 hPa (see Fig. 6c-d). The significance of these differences is slightly more consistent across the seasons for the radiosonde data than 12 hours earlier (i.e., differences at 24 hours are also significant for spring phase 1 and close to significant for autumn phase 1 and 2), whereas only the phase 1 annual average is significant for the aircraft data (see Fig. 6e-f). All of the significant differences for the radiosonde data once again meet the yellow criteria and this significance carries over to both sub-domains. On the other hand, none of the aircraft differences in this layer meet the yellow criteria and the significance of this difference does not carry over to either sub-domain. Below 300 hPa, the core with the smallest RMSE now varies with season and physics package for both radiosonde and aircraft data. Those differences that are statistically significant below 300 hPa do not exhibit any consistency between the two verification data sets.

3.1.1.2 Bias

Forecast Hour 00

The initial wind speed bias for both dynamical cores is negative at all levels (see Fig. 7a-b). The magnitude of the initial ARW bias is larger than that of the NMM above 400 hPa and below 700 hPa, with the maximum difference occurring around 200 hPa (see Fig. 7c-d). The initial wind speed bias differences in these layers are statistically significant (see Fig. 7e-f), which is also generally true for both verification sub-domains (not shown). Only bias differences in the layer above 300 hPa meet the yellow criteria.

Forecast Hour 12

The wind speed bias 12 hours into the forecast remains negative for all seasons and physics packages except at the lowest and highest levels (see Fig. 8a-b). Whether the negative bias is larger than that at the initial time depends on the level and the season. The wind speed bias for the summer season exhibits the least variability with height, whereas the wind speed bias for the winter and spring seasons exhibits the largest variability with height. The ARW tends to exhibit a larger negative bias in the layer 550-250 hPa, but the bias differences exhibit a seasonal dependence (see Fig. 8c-d). Note that the difference profiles for 12 h do not show a strong

correlation to those at the initial time. The statistical significance of the bias differences in the layer 400-250 hPa is the most consistent across phases, seasons and data type (see Fig. 8e-f), with a number of the differences in this layer meeting the yellow criteria. On the other hand, the statistical significance of the differences at these levels does not consistently carry over to both sub-domains (not shown).

The difference profiles for both the radiosonde and aircraft data suggest that the ARW tends to exhibit a smaller negative bias around 700 hPa (Fig. 8c-d), but none of these differences are statistically significant and all differences at this level meet the green criteria (Fig. 8e-f). On the other hand, a number of the differences below 700 hPa and above 250 hPa are statistically significant and meet the yellow criteria. However, the sign of the difference depends on the season at these levels (i.e., the dynamical core with the smallest magnitude wind speed bias depends on the season at these levels).

Forecast Hour 24 h

The overall characteristics of the wind speed bias 24 hours into the forecast are similar to those 12 hours earlier except the negative bias at mid-levels has decreased slightly and the negative bias at upper levels has increased slightly (see Fig. 9a-b). The strongest signal in the bias difference continues to be at upper levels (see Fig. 9c-f). In the layer 250-200 hPa, the magnitude of the ARW bias is larger than that of the NMM, which corresponds to the ARW having a larger negative bias. At 150 hPa, the magnitude of the ARW bias is smaller than that of the NMM, which corresponds to the ARW having a smaller positive bias. This trend indicates the ARW winds at upper levels tend to be weaker than those in the NMM. The statistical significance of the radiosonde differences at 200 and 250 hPa generally carries over to both sub-domains for the annual average, but not for the radiosonde seasonal averages or any of the aircraft averages (not shown). The statistical significance for the differences at 150 hPa does not carry over to both sub-domains for any of the averages (not shown). Most of the statistically significant bias differences at 24 hours also meet the yellow criteria (Fig. 9e-f).

The difference profiles for the aircraft data continue to suggest the ARW tends to exhibit a negative bias that is smaller than that for the NMM around 700 hPa, but the differences at this level for the radiosonde data are now centered on zero (Fig. 9c-d). These differences continue to meet the green criteria and are not statistically significant for either data set (Fig. 9e-f).

The difference fields in the lowest layer show more consistency across seasons and data type than 12 hours earlier, with summer being the sole outlier (Fig. 9c-d). Note that, for the most part, all the seasons except for summer have a positive bias at this level. Hence, the ARW tends to have a larger positive bias (all seasons except summer) and a smaller negative bias (summer) at this level, which would correspond to the lower level winds in the ARW tending to be stronger than that in the NMM. The statistically significant differences at this level also tend to meet the yellow criteria (Fig. 9e-f).

3.1.2 Temperature

3.1.2.1 RMSE

Forecast Hour 00

The initial temperature RMSE for both dynamical cores generally decreases slightly with height up to about 300 hPa and then increases with height above this level (see Fig. 10a-b). The aircraft RMSE is generally larger than the corresponding radiosonde RMSE and undergoes a larger

increase with height above 300 hPa. Summer tends to exhibit the smallest RMSE and winter and autumn the largest. The difference plots for both data types indicate the dynamical core with the smallest initial RMSE varies with height, with the NMM tending to have the smallest errors at lower and upper levels, and the ARW having the smallest errors at intermediate levels (see Fig. 10c-d). A number of these differences are statistically significant, with only the radiosonde differences above 250 hPa showing consistency across seasons (see Fig. 10e-f). On the other hand, all of these differences fall well within the green criteria.

Forecast Hour 12

During the first 12 hours of the forecast, the temperature RMSE increases at all levels (see Fig. 11a-b). This increase is such that the initial minimum at intermediate levels becomes more pronounced at 12 hours. Summer continues to have the smallest overall RMSE. The ARW RMSE is generally smaller than the NMM RMSE in the 400-200 hPa layer (except spring at 400 hPa), whereas the NMM RMSE tends to be smaller than the ARW RMSE in the 700-550 hPa layer (except phase 2 summer – see Fig. 11c-d). The RMSE differences in the 400-200 hPa layer tend to be statistically significant for both phase 1 and 2, with the radiosonde differences showing the strongest consistency across season. Conversely, the RMSE differences for the 700-550 hPa layer are only statistically significant for phase 1, with the exception of the winter season for the aircraft data (see Fig. 11e-f). The significance of these temperature RMSE differences does not always carry over to the West sub-domain (not shown). Although the RMSE differences for the 400-200 hPa layer are the largest and exhibit the most consistent statistical significance across seasons and phases, only the radiosonde autumn differences at 200 hPa meet the yellow criteria.

Forecast Hour 24

During the second 12 hours of the forecast, the temperature RMSE continues to increase at all levels, while maintaining a minimum at intermediate levels (see Fig. 12a-b). The ARW RMSE continues to be smaller than the NMM RMSE in a layer between 400-250 hPa (see Fig. 12c-d) and the differences once again tend to be statistically significant for both phase 1 and 2 (see Fig. 12e-f). This statistical significance tends to carry over to both sub-domains (not shown), but the magnitude of the differences in this layer decreases over the second 12 hours of the forecast. Note that none of the differences in this layer meet the yellow criteria at this forecast time. The NMM RMSE for phase 1 still tends to be smaller than the ARW RMSE in the layer 700-550 hPa and this difference is statistically significant for most seasons, but this signal definitely does not carry over to phase 2 at this forecast time. The biggest change in the RMSE differences occurs at low levels. The ARW RMSE is now smaller than the NMM RMSE for all seasons and phases except summer for the aircraft data. The magnitude of the differences in this layer has generally increased, and the number of differences that are statistically significant has increased. When considering the verification sub-domains, the statistical significance of these low-level RMSE differences are more likely to carry over to the West/East sub-domain than the East/West sub-domain for the radiosonde/aircraft data (not shown).

3.1.2.2 Bias

Forecast Hour 00

The initial temperature bias for both dynamical cores is, for the most part, negative below 300 hPa and positive above 300 hPa, with the exception that the sign of the bias at the lowest levels

depends on the observational data type (see Fig. 13a-b). Although the differences between the initial biases of the two dynamical cores tend to be statistically significant, the differences are rather small (i.e., all differences fall well within the green criteria) and tend to lack sign consistency across season and observational data type (see Figs. 13c-f).

Forecast Hour 12

During the first 12 hours of the forecast, the positive temperature bias at upper levels increases for both dynamical cores (see Fig. 14a-b). This behavior is consistent across all seasons and physics packages. On the other hand, the evolution of the negative temperature bias at 700 hPa shows more dependence on the season and physics package. The ARW temperature bias generally undergoes a larger increase than the NMM and the increase is generally larger for both dynamical cores when run with phase 1 physics. The magnitude of the ARW temperature bias tends to larger than the NMM bias below 500 hPa (one exception is summer, for which the ARW bias is only larger than the NMM bias around 700 hPa), whereas it tends to be smaller than the NMM bias above 300 hPa (see Fig. 14c-d). This transition in the bias difference stems from the fact that the ARW tends to be colder than the NMM at all levels. These temperature bias differences tend to be statistically significant for most of the seasons at all levels except the lowest (see Fig. 14e-f). The layers for which the bias differences exhibit the most consistency across seasons (including both sign and statistical significance) are 300-200 hPa (ARW bias smaller than NMM) and 800-500 hPa (NMM bias smaller than ARW). The differences in the lower layer tend to fall in the yellow category, whereas the differences in the upper layer tend to fall in the yellow to red categories. The statistical significance of the differences in these two layers generally carries over to the verification sub-domains (not shown).

Forecast Hour 24

The temperature bias profiles at 24 hours are very similar to those at 12 hours, with the negative bias at 700 hPa increasing slightly over the second 12 hours of the forecast (see Fig. 15a-b). The difference profiles for the temperature biases are also very similar to that at 12 hours, except that the differences in the 800-500 hPa layer are slightly larger and summer phase 1 now follows the profiles for the other seasons more closely than the profile for summer phase 2 (see Fig. 15c-d). These temperature bias differences once again tend to be statistically significant for most of the seasons at all levels except the lowest (see Fig. 15e-f), with the layers exhibiting the most consistency across season once again being 300-200 hPa (ARW bias smaller than NMM) and 700-500 hPa (NMM bias smaller than ARW). On the other hand, more of the differences in the lower layer now fall in the red category, whereas the differences in the upper layer tend to fall more in the yellow category. The statistical significance of the differences for the lower layer generally carries over into the verification sub-domains, whereas the significance of differences for the upper layer is less likely to carry over to both sub-domains for the seasonal averages (not shown).

3.1.3 Relative Humidity

3.1.3.1 RMSE

Forecast Hour 00

The relative humidity RMSE at the initial time increases slightly with height for both dynamical cores, with the summer season exhibiting the largest overall variability with height (see Fig. 16a). The ARW RMSE is generally larger than the NMM RMSE, with the difference being the

largest at 500 hPa (see Fig. 16c). Only the RMSE differences at 500 hPa are statistically significant across all seasons (see Fig. 16e). All of the initial RMSE differences fall in the green category.

Forecast Hour 12

During the first 12 hours of the forecast, the relative humidity RMSE increases at all levels such that the vertical distribution of the errors remains basically the same as that at the initial time (see Fig. 17a). On the other hand, the differences between the two dynamical cores have undergone a marked change. The ARW RMSE now tends to be smaller than the NMM RMSE for the lowest two levels (see Fig. 17c), but these differences are only statistically significant for phase 2 (see Fig. 17e). The NMM RMSE tends to be smaller than the ARW RMSE at 500 hPa, but only the difference for summer phase 2, for which the ARW RMSE is actually smaller than the NMM RMSE, is statistically significant. Most of the statistically significant RMSE differences at this time also meet the yellow criteria.

Forecast Hour 24

During the second 12 hours of the forecast, the relative humidity RMSE continues to increase at all levels, but the size of this increase is smaller than the first 12 hours (see Fig. 18a). The RMSE differences at 24 hours exhibit more variability across seasons and physics packages than those at 12 hours, whereas the differences undergo a general shift such that the ARW RMSE shows a stronger tendency to be smaller than the NMM RMSE at all three levels (see Fig. 18c). On the other hand, these differences do not show consistent statistical significance across seasons and physics packages (see Fig. 18e).

3.1.3.2 Bias

Forecast Hour 00

The initial relative humidity biases for both cores hover around zero at 850 hPa, progress toward small positive biases at 700 hPa for all but the summer season, and finally transition to all positive biases at 500 hPa (see Fig. 16b). The magnitude of the NMM bias at 850 hPa is larger than that for the ARW, except for the autumn season, whereas the magnitude of the ARW bias at 700 and 500 hPa is larger than that for the NMM, except for the summer season at 700 hPa (see Fig. 16d). Most of these differences are statistically significant, but only those at 500 hPa tend to meet the yellow criteria (see Fig. 16f). The statistical significance of the initial bias differences carries over to the sub-domains at 500 hPa and for the most part 700 hPa, whereas the behavior of the statistics at 850 hPa exhibit less consistency across the sub-domains (not shown).

Forecast Hour 12

The relative humidity biases at 12 hours are negative at 850 hPa and transition to positive at 500 hPa, except for the summer season (see Fig. 17b). For most seasons and levels, the magnitude of the ARW bias is larger than that for the NMM (see Fig. 17d). Although not immediately obvious from the difference profiles, closer inspection reveals that the ARW relative humidity tends to be larger than that for the NMM at 700 and 500 hPa. Note that the higher ARW relative humidity does not necessarily translate to the ARW being moister than the NMM because the tendency for the ARW to be colder than the NMM (see temperature bias discussion above) could, by itself, lead to the ARW relative humidity being higher than that of the NMM. The differences at 850 hPa are only statistically significant for the annual and winter summer averages (see Fig. 17f), whereas the differences at 700 and 500 hPa are statistically significant

for a broader range of seasons, but the statistically significant differences are not all of the same sign. Conversely, most of the statistically significant bias differences for relative humidity meet the red criteria.

Forecast Hour 24

During the second 12 hours of the forecast, the magnitude of the relative humidity bias increases slightly for most seasons and levels, while maintaining a vertical distribution similar to that found at 12 hours. The magnitude of the ARW bias is, once again, generally larger than that for the NMM, with the exception of the phase 2 summer and annual averages at 700 and 500 hPa. These exceptions are once again consistent with a tendency for ARW relative humidity to be higher than that of the NMM at 700 and 500 hPa. The 850 hPa differences are now only statistically significant for phase 2 annual and summer averages, whereas the 700 and 500 hPa differences are, once again, significant for a broad range of seasons for both phases, but not all of the same sign.

3.2 Precipitation

Both QPF verification measures (ETS and bias) are positive quantities. The higher the ETS, the more skillful the forecast, whereas a bias of one is the most skillful. All ETS and bias differences discussed below were computed by subtracting the NMM verification measure from its ARW counterpart. Hence, a positive ETS difference means the ARW forecasts have more skill than the NMM forecasts, and a negative ETS difference means the NMM forecasts have more skill than the ARW forecasts. Interpreting the bias differences is, once again, slightly more complicated. Positive differences for biases greater than one indicate the NMM forecasts have more skill than the ARW forecasts, whereas positive differences for biases less than one indicate the ARW forecasts have more skill than the NMM forecasts. A bold dotted line has been added to the bias differences plots indicating the transition from bias greater than one to less than one to facilitate interpreting the results.

3.2.1 Equitable Threat Score

The Equitable Threat Scores (ETS) for both dynamical cores and both physics packages are shown in Fig. 19a. These scores indicate the skill of the forecasts for all four configurations is lower for higher thresholds, as would be expected. Note that the skill of the forecasts at the lower thresholds show a stronger dependence on the physics suite than the dynamical core, with the phase 2 physics suite showing less skill than the phase 1 physics suite. For phase 1, the ARW shows less skill than the NMM for all but the highest threshold, but only the difference for the lowest threshold is statistically significant (see Fig. 19c). For phase 2, the ARW shows more skill for all but one of the thresholds (see Fig. 19e). Only the differences for the lowest three thresholds are statistically significant for phase 2, all of which indicate the ARW has more skill that the NMM. Hence, the dynamical core with the higher skill is opposite that for phase 1. In addition, all of the ETS differences for both phase 1 and 2 fall well within the green category.

3.2.2 Bias

The bias for both dynamical cores is greater than one for the lower thresholds and decreases to less than one for the higher thresholds (see Fig. 19b). Hence, the forecasts overestimate the areal coverage of the precipitation at the lower thresholds and underestimate the areal coverage at the higher thresholds. Note that the bias also shows a stronger dependence on the physics suite than the dynamical core, with phase 2 physics having the higher bias for all thresholds. For phase 1,

the ARW bias is greater than that of the NMM for all but the highest thresholds. The transition from bias greater than one to less than one is such that the NMM has less overestimation of the areal coverage at the lower thresholds, more underestimation of the areal coverage at intermediate thresholds, and less underestimation at the highest thresholds (see Fig. 19d). For phase 2, the ARW bias is, once again, greater than that of the NMM for the all but the highest thresholds. The bias transition for phase 2 is such that the NMM produces a better estimate of the areal coverage for all thresholds (see Fig. 19f). Only the bias differences for the lowest thresholds are statistically significant, all of which fall in the green category.

4. Bulk Timing Statistics

Bulk timing information was saved for each forecast cycle (i.e., the time from model start to model end, including all time for I/O operations). This timing information was used to compute a ratio of the ARW runtime to the NMM runtime for each forecast cycle and then an average ratio was computed for each physics suite. It should be noted that I/O timing for the two dynamical cores could have a significant impact on these bulk timing statistics since the number of fields written to the history files is not the same for the two dynamical cores. In addition, for the phase 1 runs, the number of statements written to the log files was not equivalent for the two dynamical cores because the debug level controlled by the namelist was set differently for the two cores. And finally, the mp_zero_out option was not working properly for the phase 2 runs, which ended up having a larger impact on the runtime for the ARW phase 2 runs because the NMM has a separate minimum value control in its module_physics_calls.f routine that gave the NMM core a computational advantage. With these caveats in mind, the ARW/NMM timing ratios for the retrospective runs were 0.983 for phase 1 and 1.278 for phase 2. The DTC plans to address the need to more precise timing statistics for the two dynamical cores in the coming months by producing runs that provide a breakdown of the timing required by each basic routine.

Table 1: Proposed threshold criteria. Differences that fall in the green category are considered insignificant. Differences that fall in the yellow category are considered to be of concern. Differences that fall in the red category are considered to be of serious concern.

	Green	Yellow	Red
Winds	< 0.10 ms^{-1}	0.10 to 0.25 ms^{-1}	> 0.25 ms^{-1}
Temperature	< 0.1 K	0.1 to 0.2 K	> 0.2 K
Relative Humidity	< 0.5%	0.5 to 1.0 %	> 1.0%
Equitable Threat Score	< 0.03	0.03 to 0.05	>0.05
Precipitation Bias	< 0.1	0.1 to 0.25	>0.25

References

Benjamin, S. G., B. E. Schwartz, E. J. Szoke, and S. E. Koch, 2004: The value of wind profiler data in U. S. weather forecasting. *Bull. Amer. Meteor. Soc.*, **85**, 1871-1886.

Bernardet, L. R., L. Nance, H.-Y. Chuang, A. Loughe, M. Demirtas, S. Koch, and R. Gall, 2005: The Developmental Testbed Center Winter Forecast Experiment. Preprint, 21st Conference on Weather and Forecasting / 17th Conference on Numerical Weather Prediction, 1-5 August 2005, Washington, D.C. American Meteorology Society.

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.

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

Janjic, Z. I., 2003a: A nonhydrostatic model based on a new approach. *Meteorology and Atmospheric Physics*, **82**, 271-285.

Janjic, Z. I., 2003b: The NCEP WRF core and further development of its physical package. 5th International SRNWP Workshop on Non-Hydrostatic Modeling, Bad Orb, Germany, 27-29 October 2003.

Kain, J., S. Weiss, G. Carbin, M. Baldwin, D. Bright, J. Hart, and J. Levit, 2005: Comparisons of different WRF configurations in a severe weather forecasting environment: The 2005 SPC/NSSL Spring Program. 6th WRF / 15th MM5 Users' Workshop, 27-30 June 2005, Boulder, CO.

Lin, Y., M. Baldwin, G. DiMego, K. Mitchell and E. Rogers, 1999: Precipitation forecasts and verifications of the NCEP Eta model. AGU Fall meeting, Dec 13-19, 1999, San Francisco, CA, paper H51C-03.

Reynolds, R. W., and T. M. Smith, 1994: Improved global sea surface temperature analyses using optimum interpolation. J. Climate, 7, 929-948.

Seaman, N., R. Gall, L. Nance, S. Koch, L. Bernardet, G. DiMego, J. Powers, and F. Olsen, 2004: The WRF Process: Streamlining the transition of new science from research into operations. 5th WRF / 14th MM5 Users' Workshop, 22-25 June 2004, Boulder, CO.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF Version 2. NCAR Tech. Note, NCAR/TN-468+STR, 88 pp. [Available from UCAR Communications, P. B. Box 3000, Boulder, CO 80307].

Weatherhead, E. C., G. Noonan, T. Fowler, L. Bernardet, L. Nance, and S. Koch, 2006: Statistical Analysis of Differences in RH, Temperature, Winds, and Precipitation between the ARW Core and NMM Core.

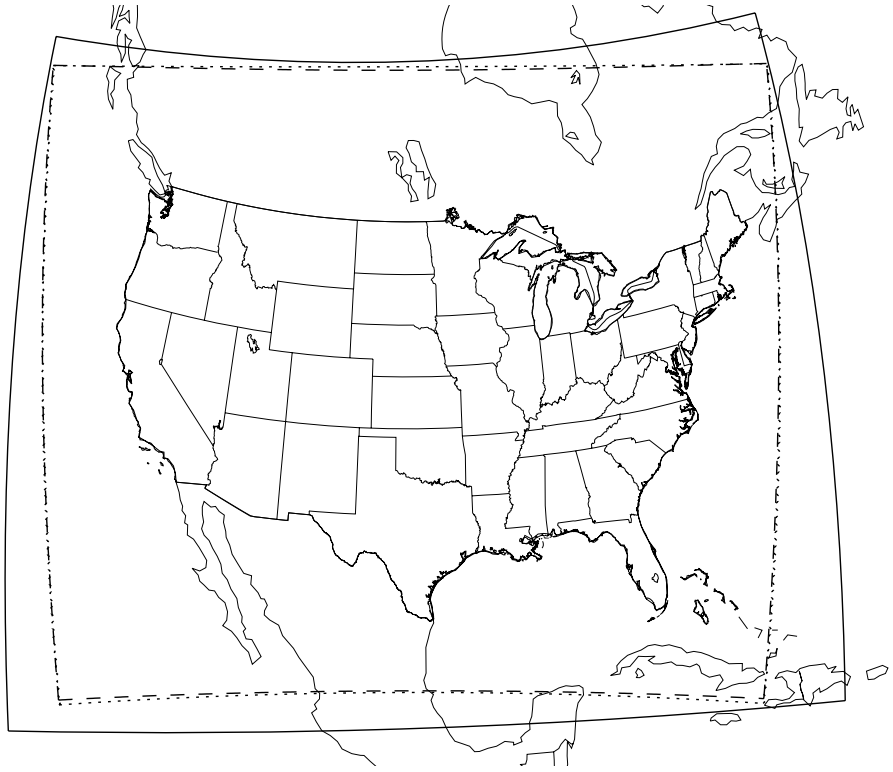

Figure 1: Map showing the boundaries of the computational domains used for the ARW (dashed line) and the NMM (dotted). The solid line shows the boundaries of the domain for the RUC13, which was used to initialized the ARW and NMM forecasts.

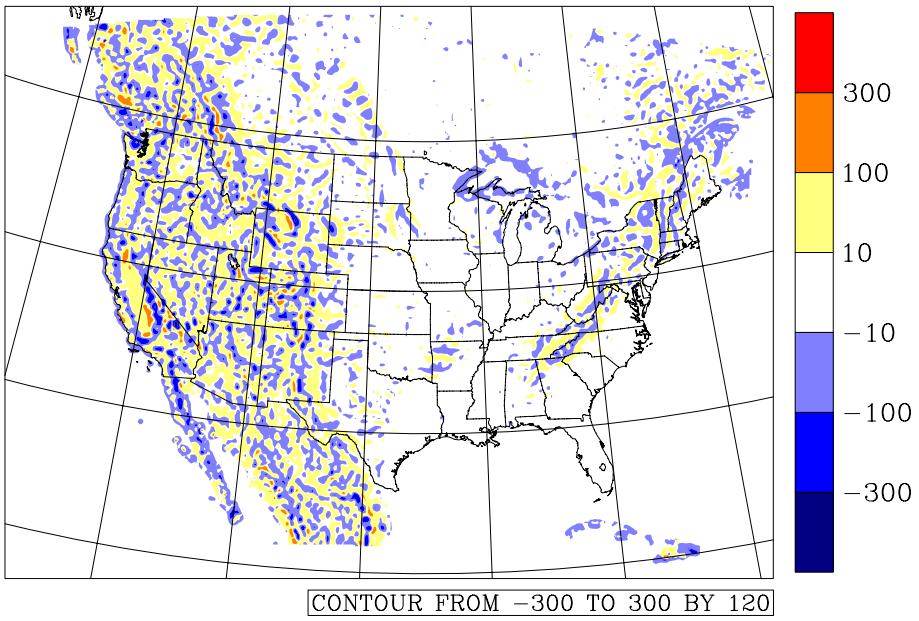

Figure 2: Contour plot showing the difference between the terrain field used in the ARW and NMM forecasts.

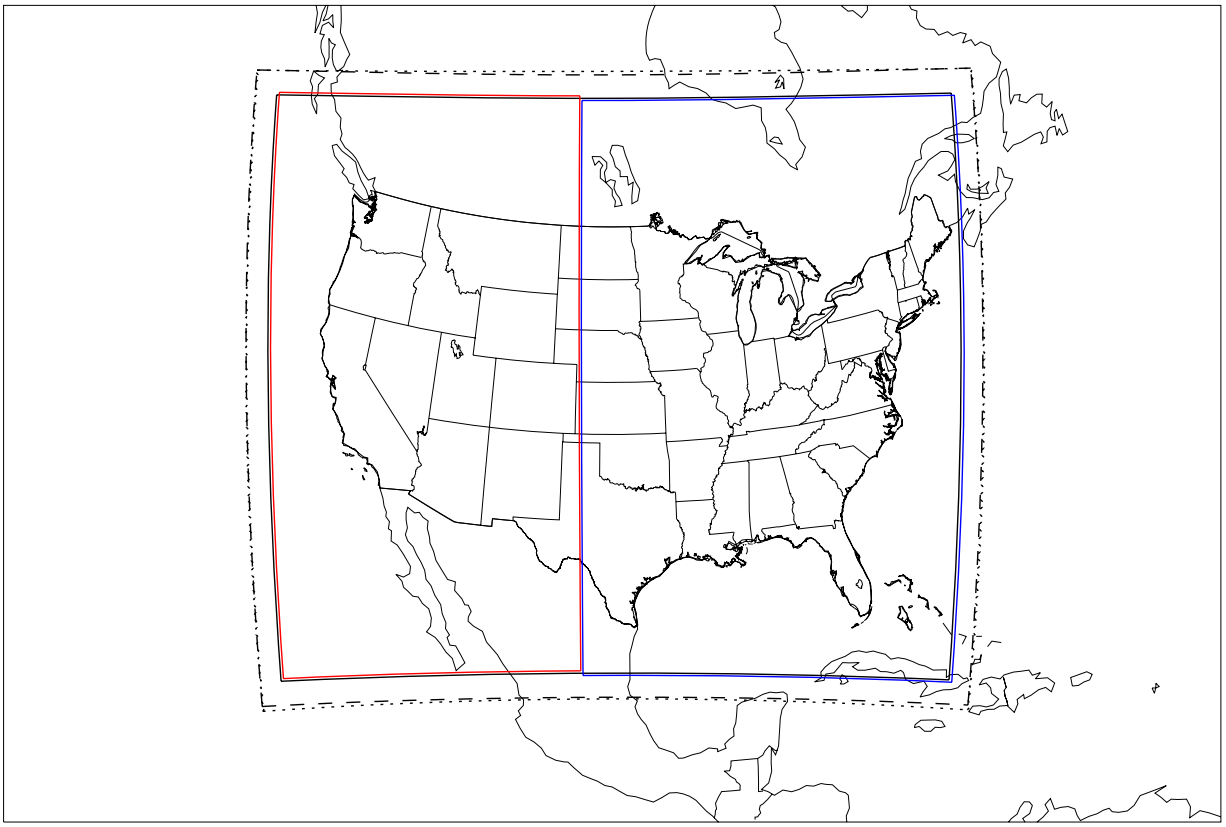

Figure 3: Map showing the boundaries of the verification domains: CONUS (solid black), West (solid red), and East (solid blue). The computational domains used for the ARW (dashed line) and the NMM (dotted) are shown for reference.

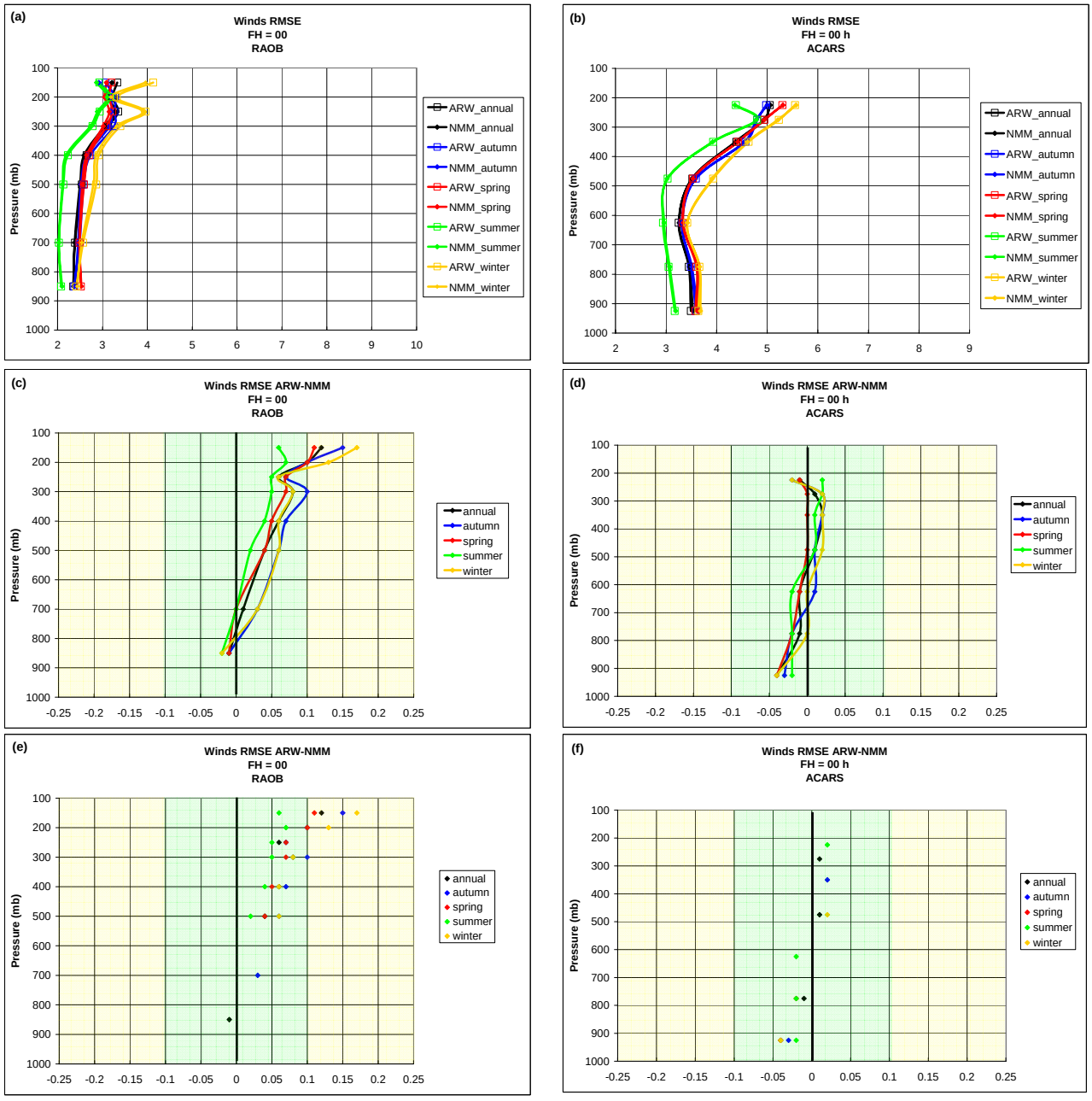

Figure 4: Vector wind RMSE (m s^{-1}) for the initial time (FH = 00 h): RMSE profiles for the radiosonde (a) and aircraft (b) data sets, RMSE difference profiles for the radiosonde (c) and aircraft (d) data sets and RMSE differences that are statistically significant for the radiosonde (e) and aircraft (f) data sets. All differences are ARW-NMM. The shading in the lower four panels corresponds to the threshold criteria summarized in Table 1.

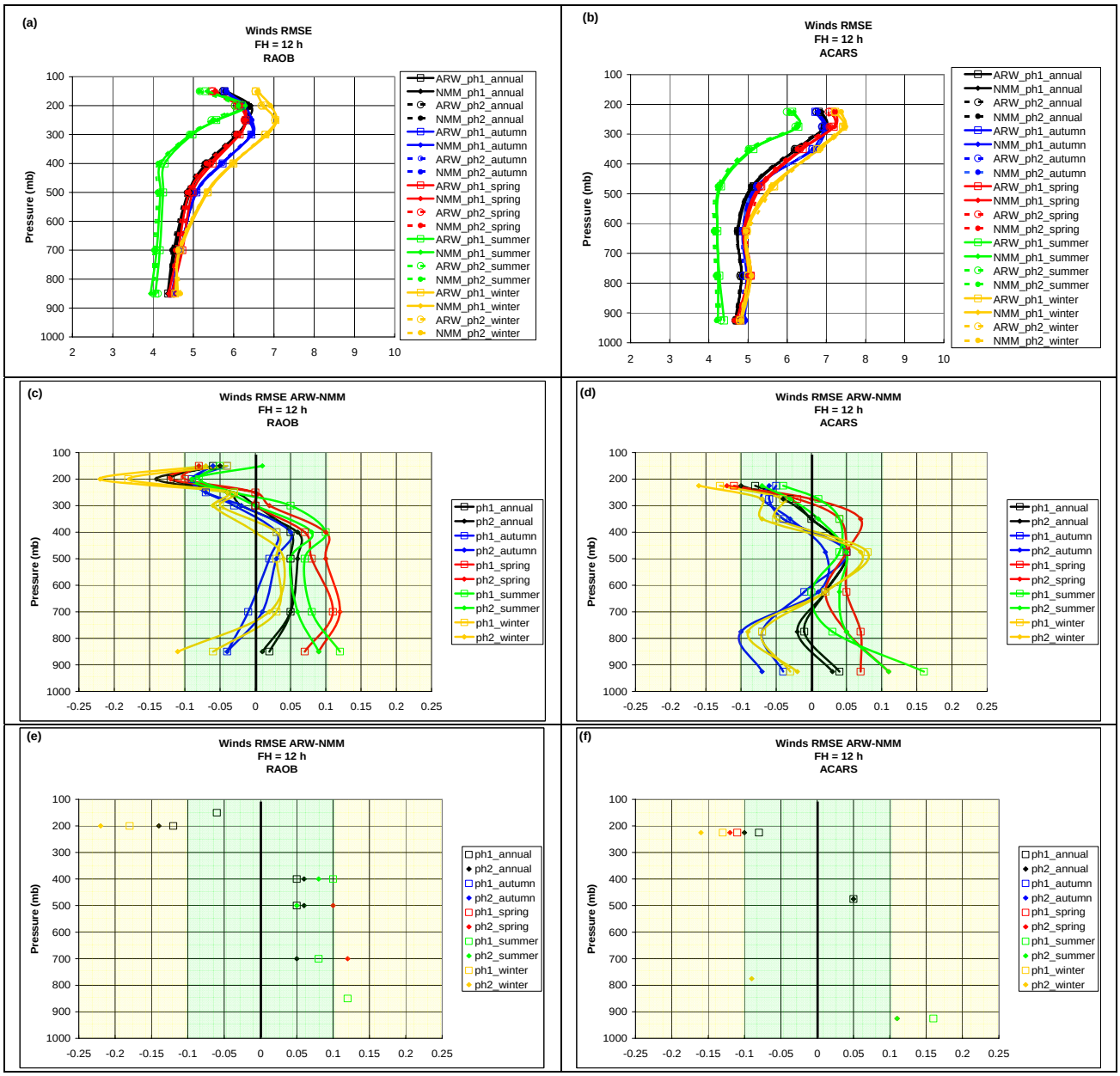

Figure 5: Same as 4 except for the vector wind RMSE for forecast hour 12 (FH = 12 h).

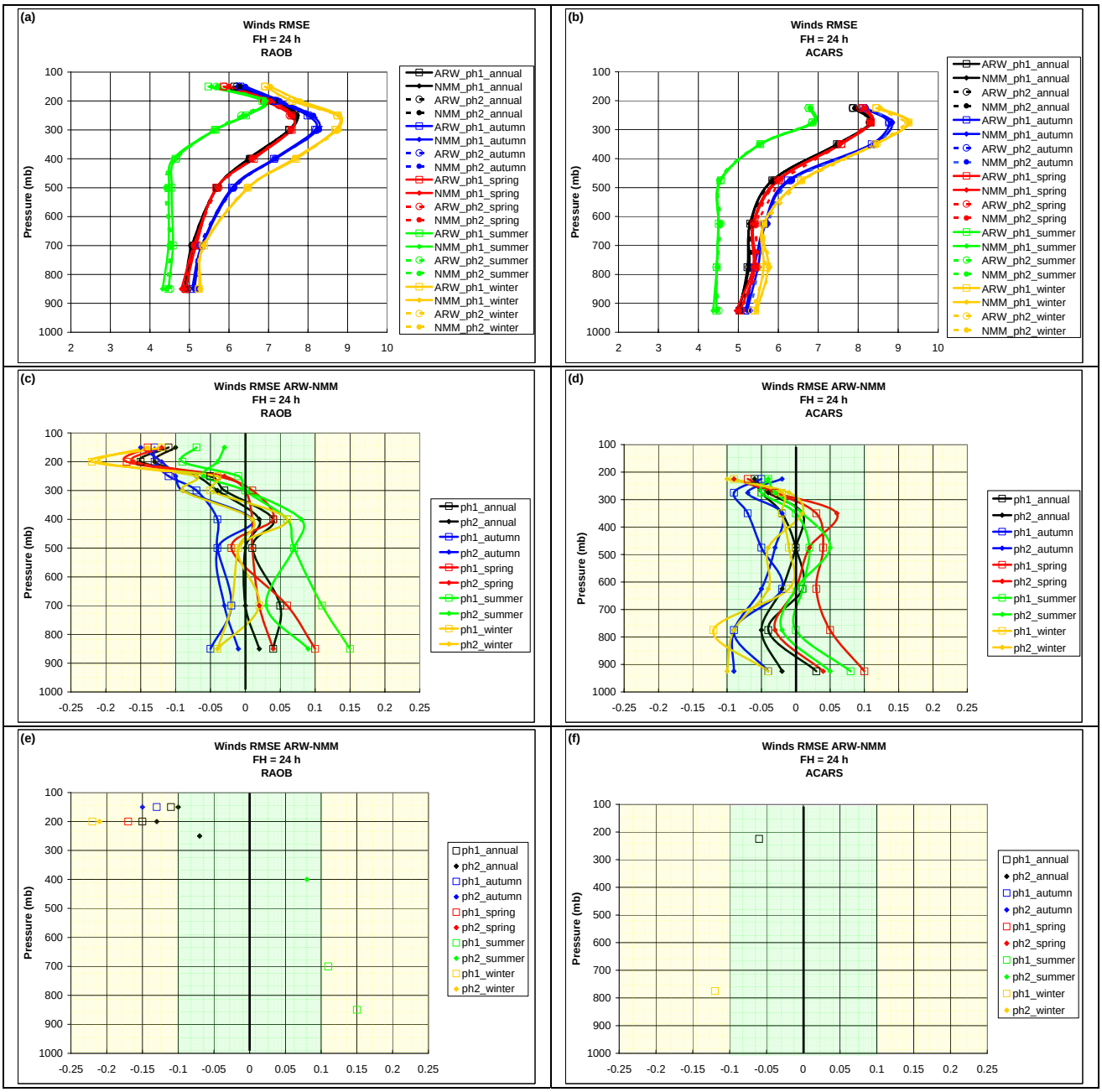

Figure 6: Same as 4 except for the vector wind RMSE (m s^{-1}) for forecast hour 24 (FH = 24 h).

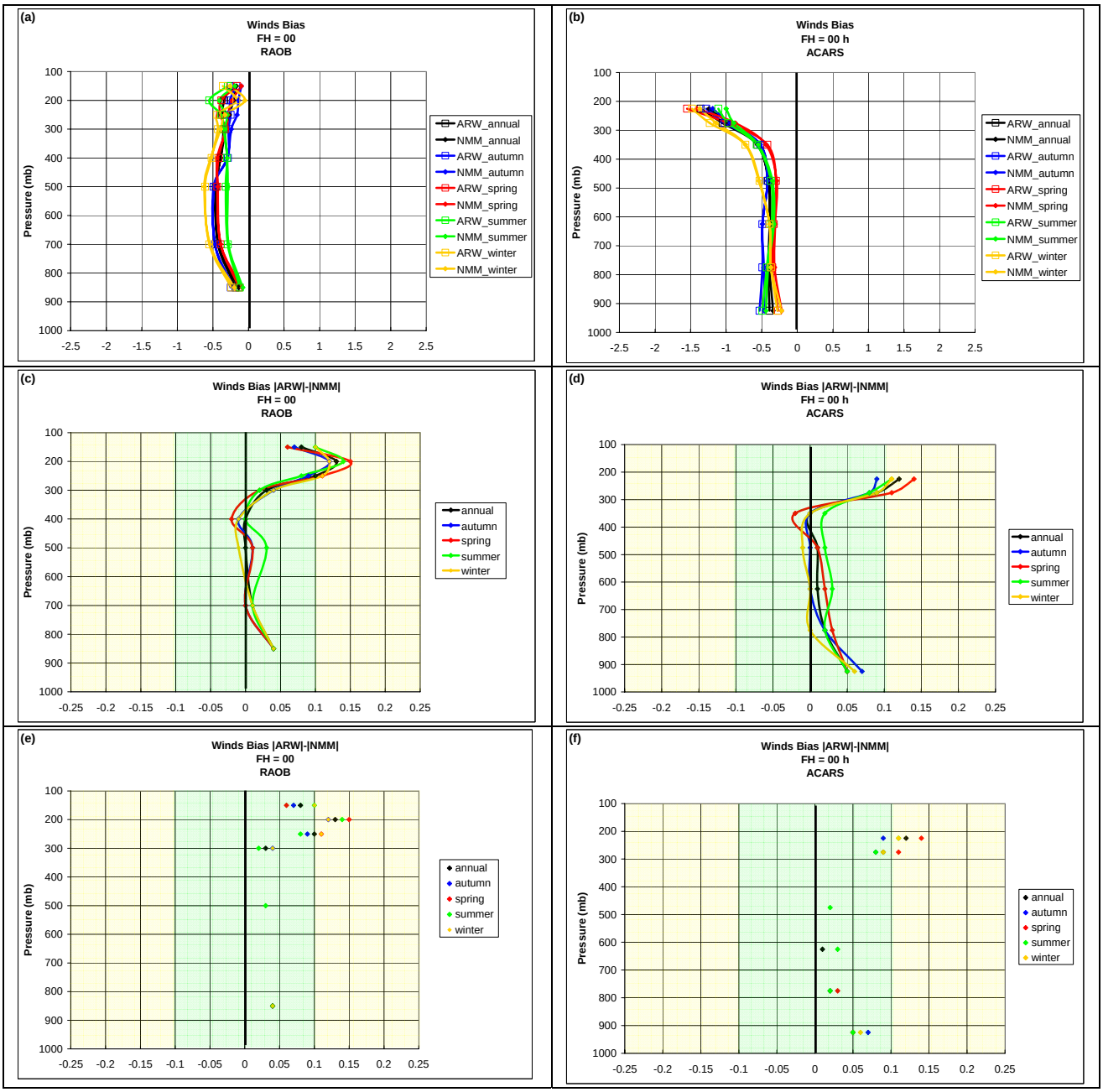

Figure 7: Same as 4 except for the wind speed bias (m s^{-1}) for the initial time (FH = 00 h). All differences are computed using the absolute value of the average bias.

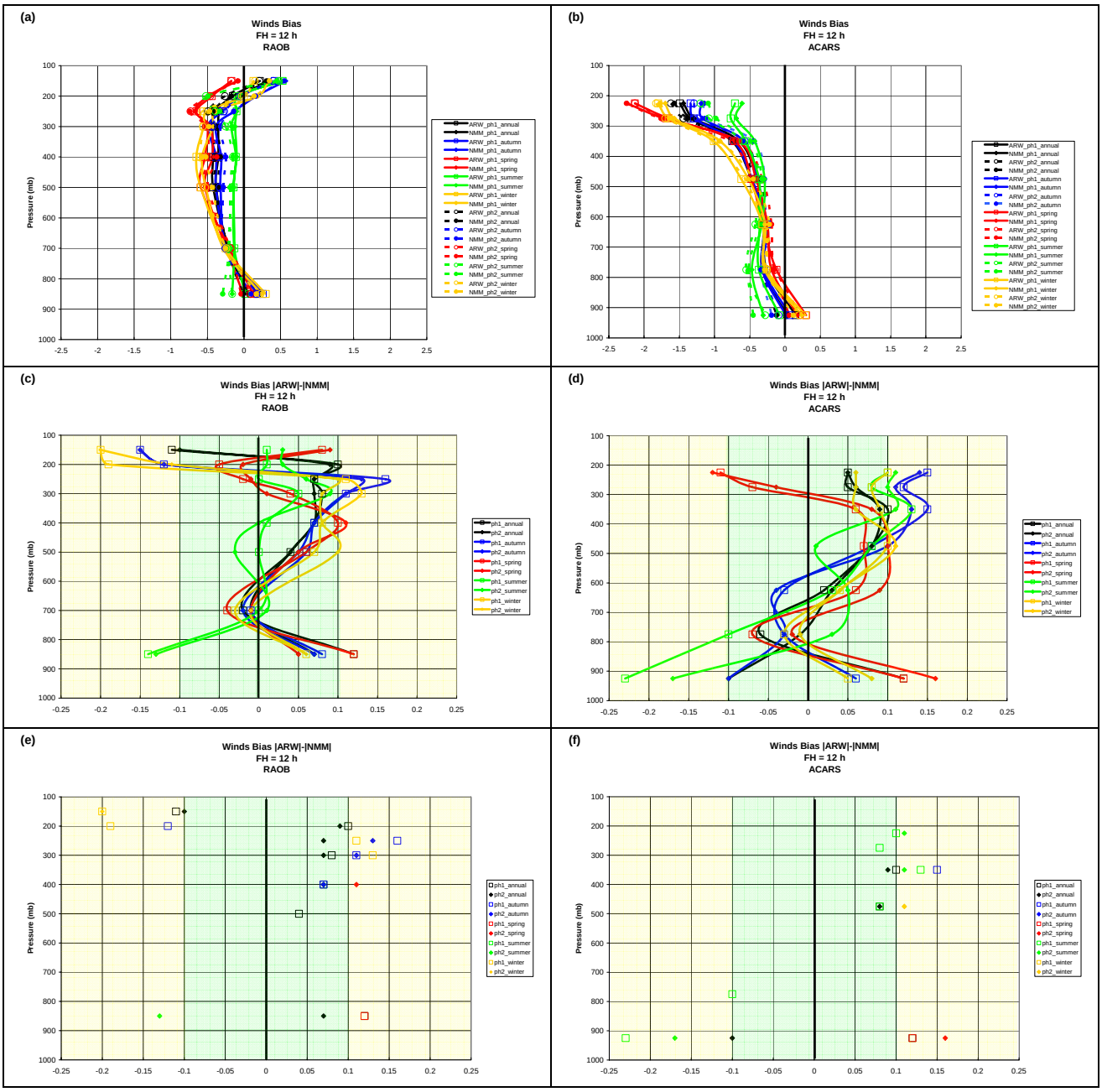

Figure 8: Same as 4 except for the wind speed bias (m s^{-1}) for forecast hour 12 (FH = 12 h). All differences are computed using the absolute value of the average bias.

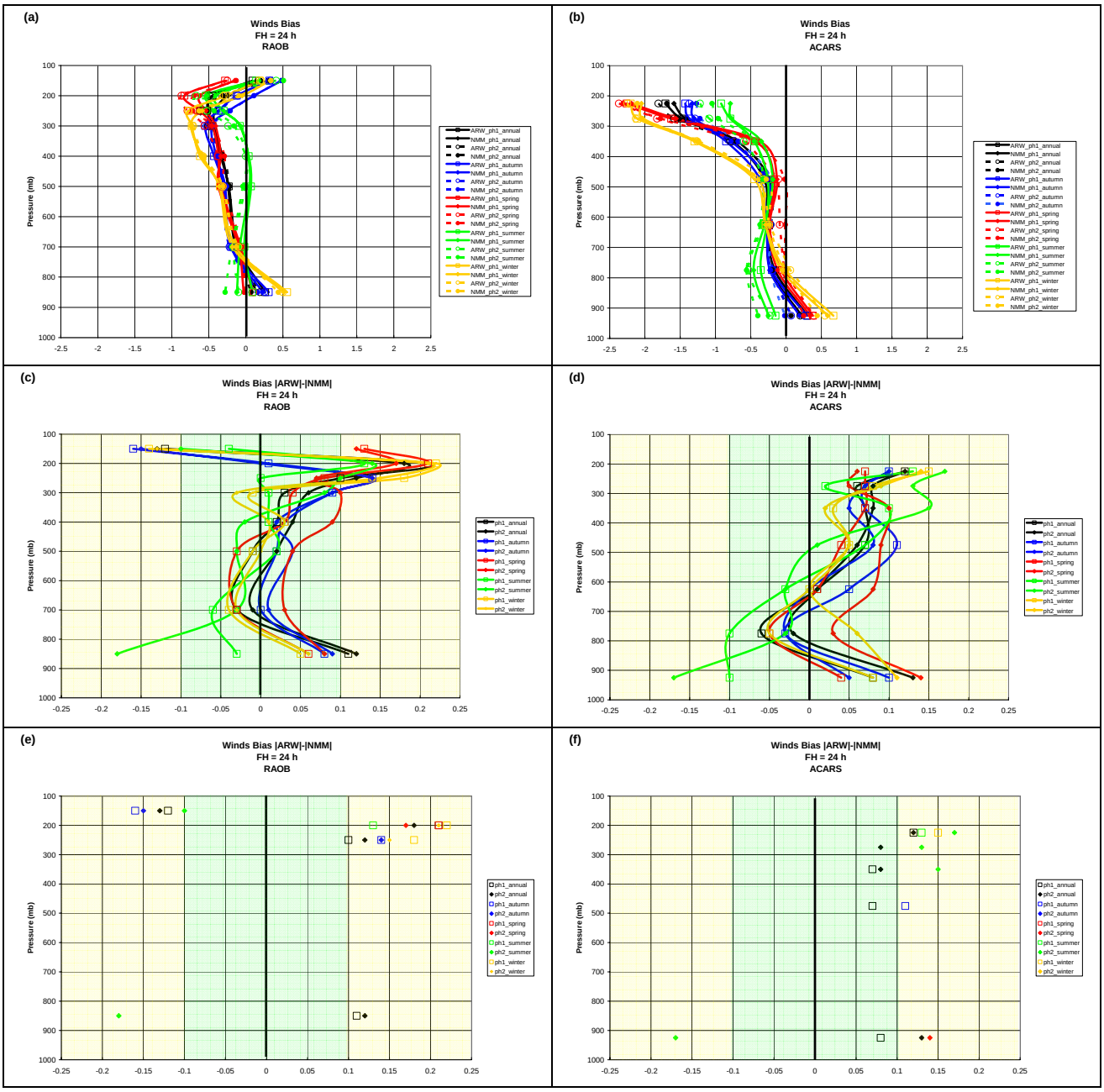

Figure 9: Same as 4 except for the wind speed bias (m s^{-1}) for forecast hour 24 (FH = 24 h). All differences are computed using the absolute value of the average bias.

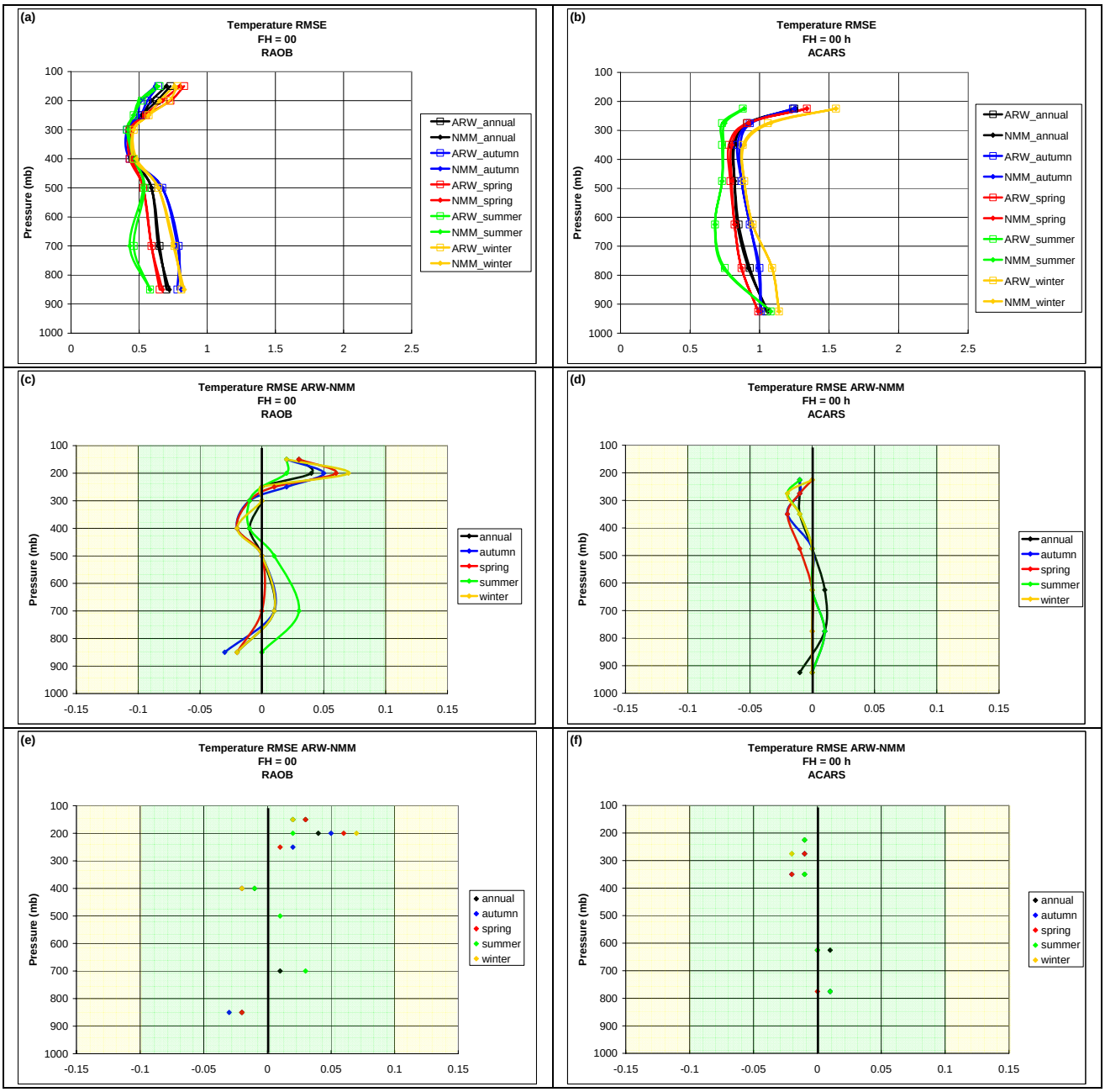

Figure 10: Same as 4 except for the temperature RMSE (K) for the initial time (FH = 00 h).

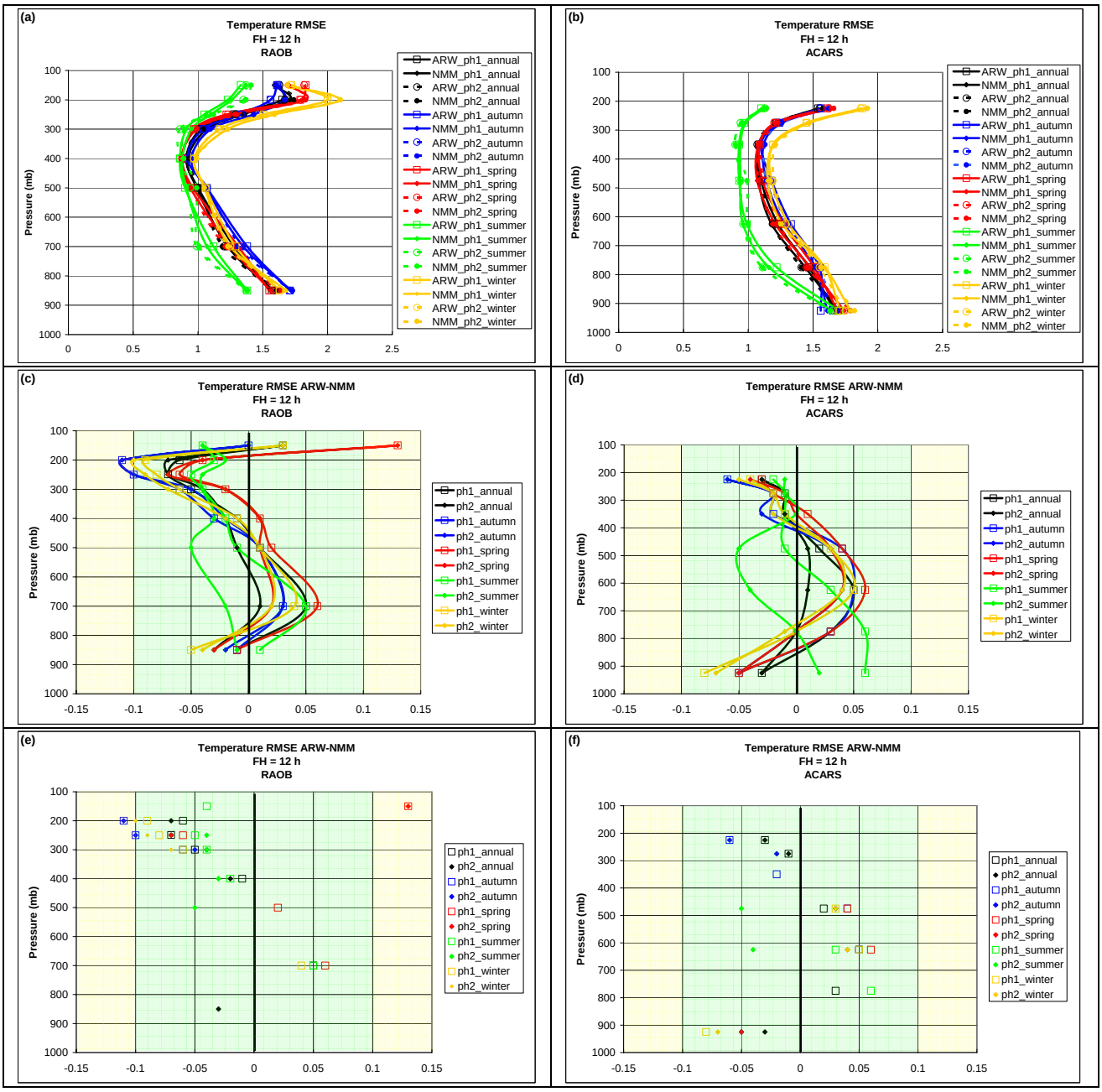

Figure 11: Same as 4 except for the temperature RMSE (K) for forecast hour 12 (FH = 12 h).

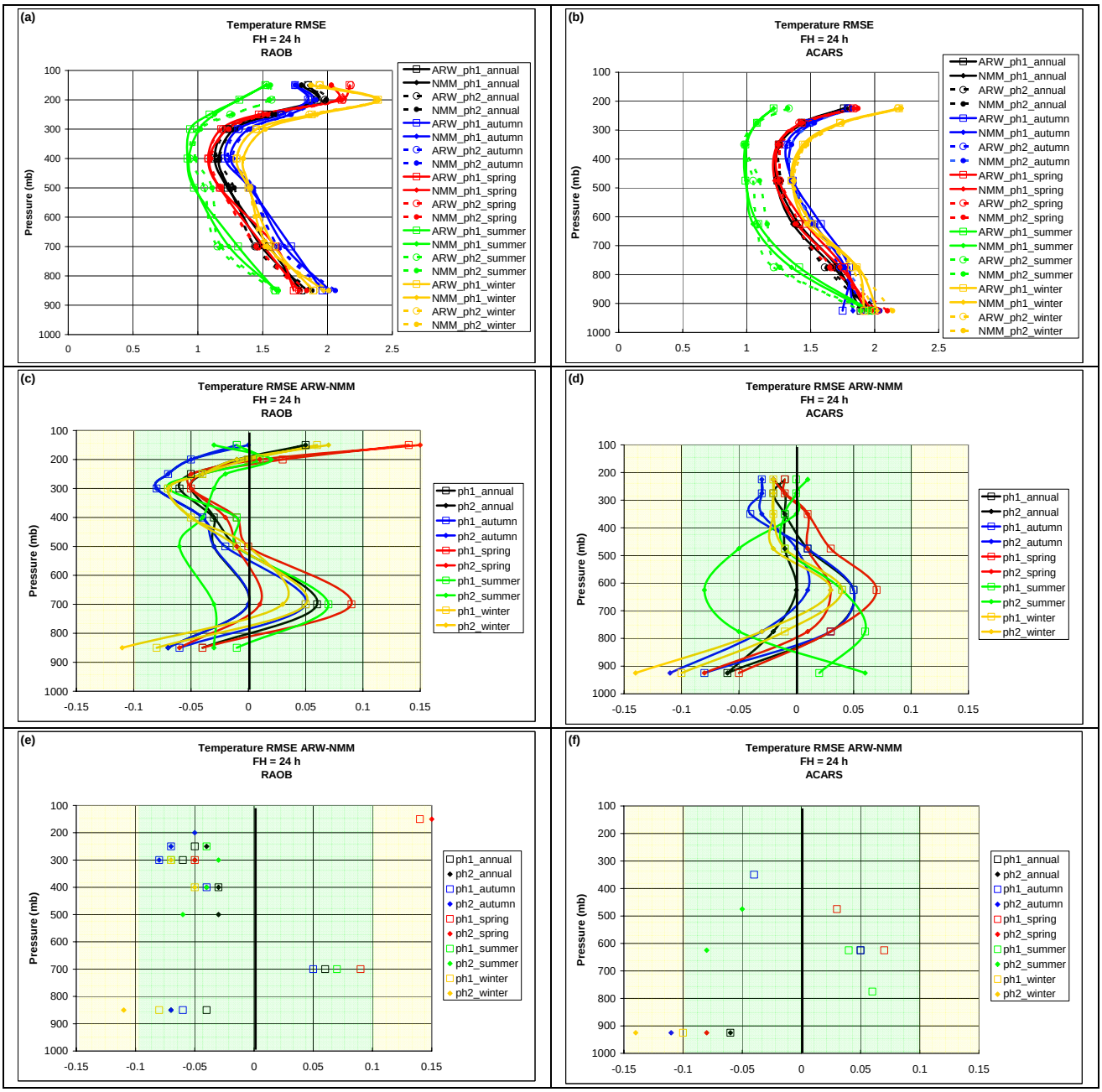

Figure 12: Same as 4 except for the temperature RMSE (K) for forecast hour 24 (FH = 24 h).

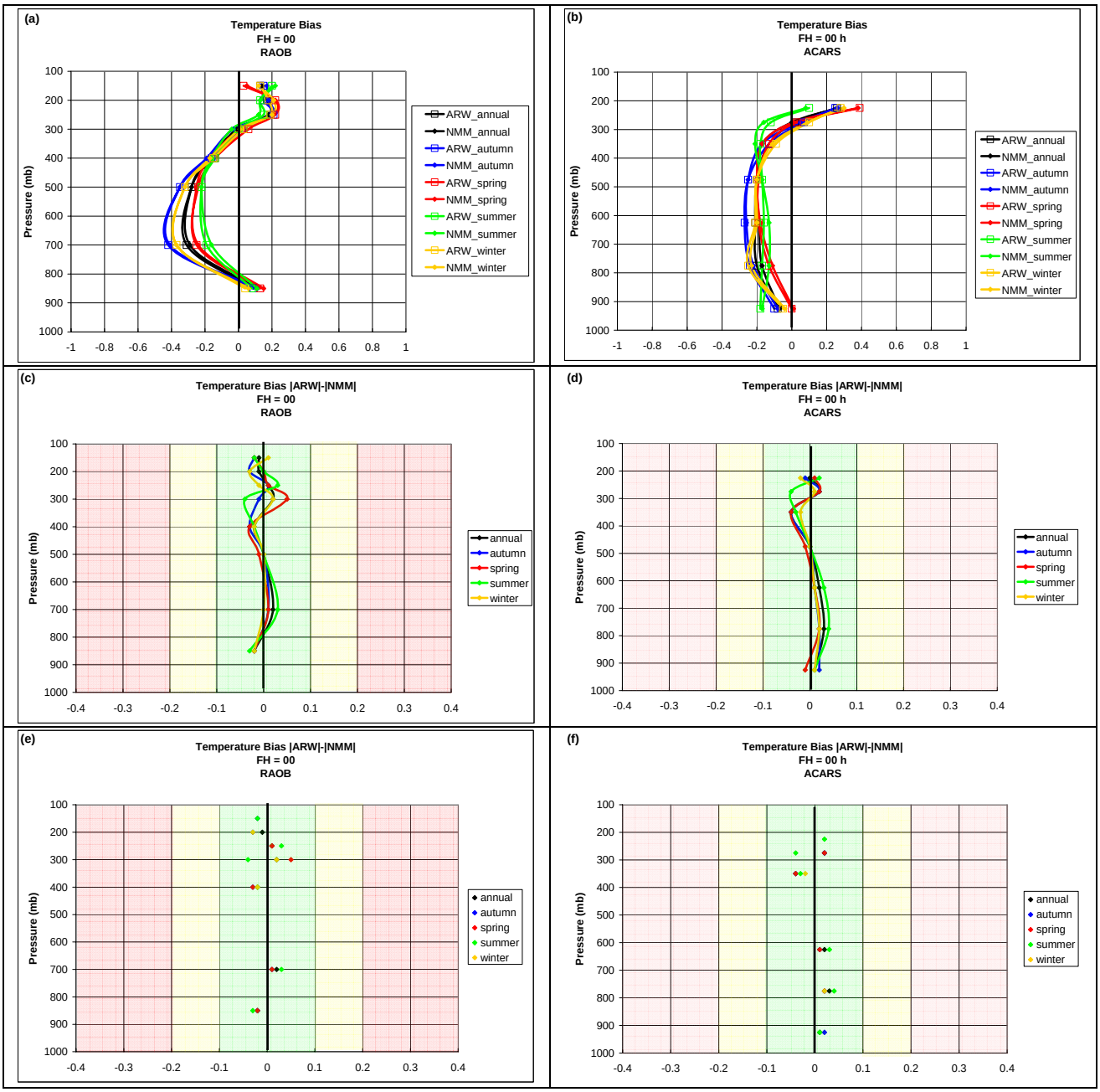

Figure 13: Same as 4 except for the temperature bias (K) for the initial time (FH = 00 h).). All differences are computed using the absolute value of the average bias.

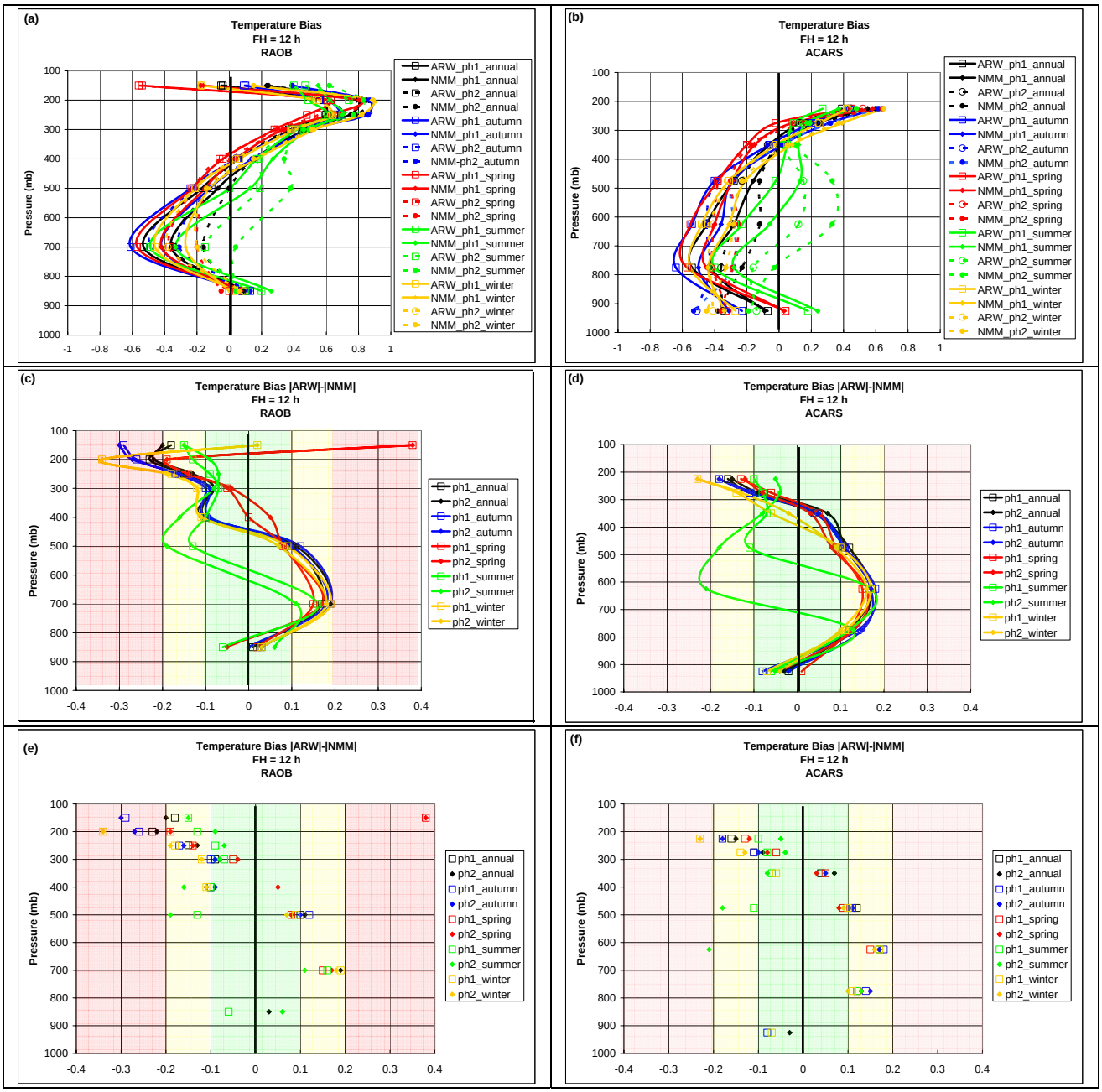

Figure 14: Same as 4 except for the temperature bias (K) for forecast hour 12 (FH = 12 h).). All differences are computed using the absolute value of the average bias.

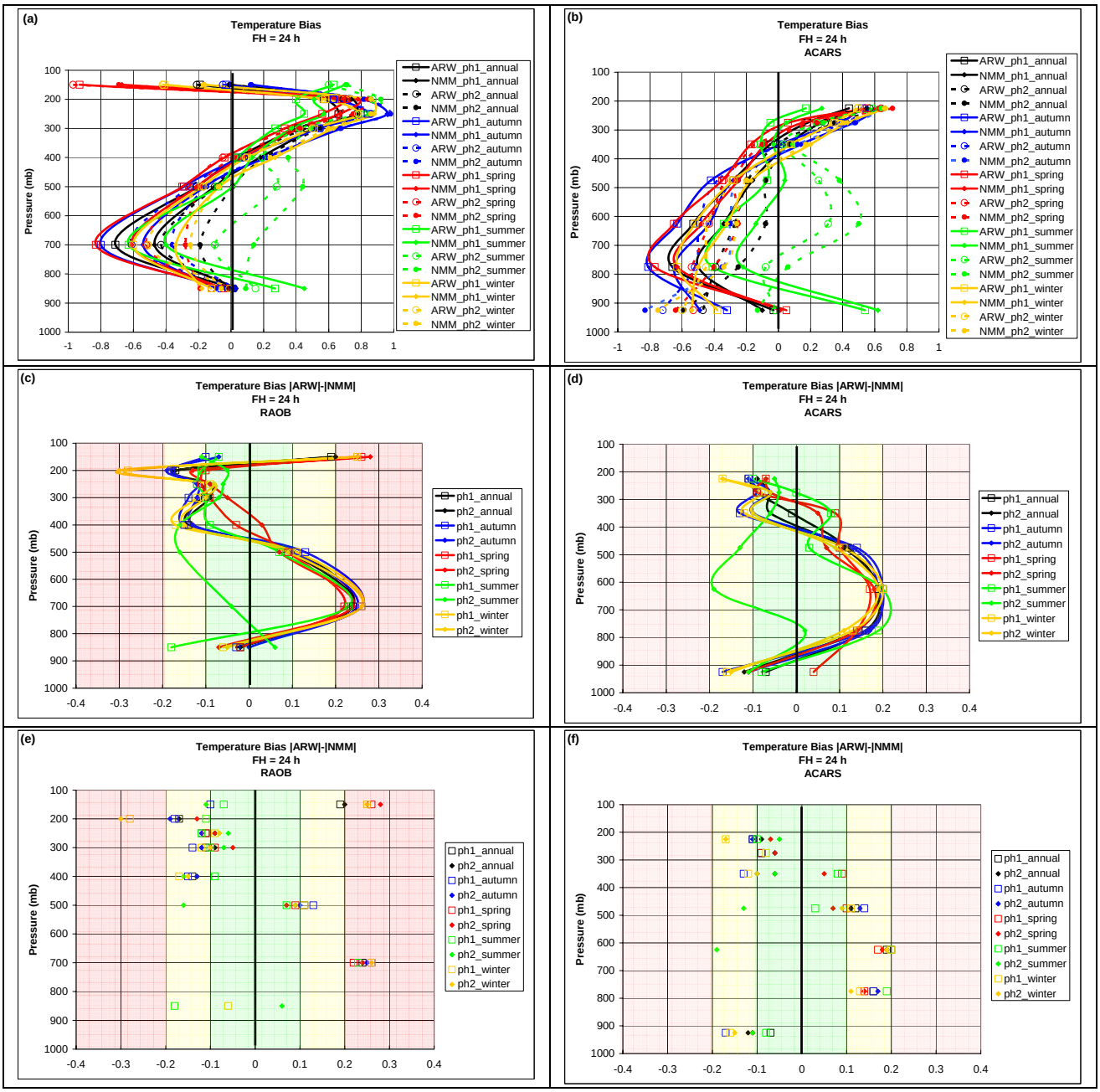

Figure 15: Same as 4 except for the temperature bias (K) for forecast hour 24 (FH = 24 h).). All differences are computed using the absolute value of the average bias.

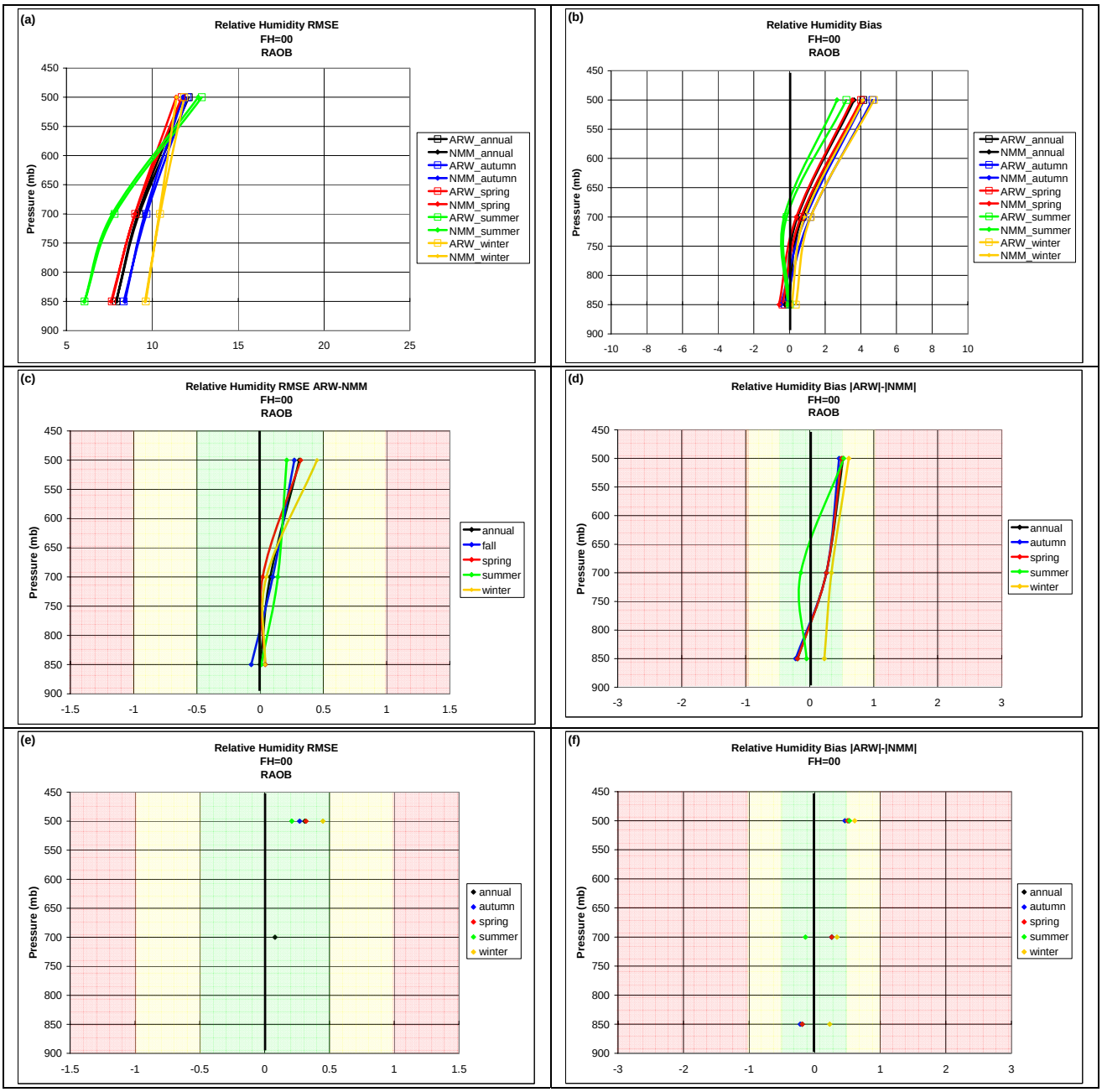

Figure 16: Relative humidity verification statistics (%) for the initial time (FH = 00 h): (a) RMSE and (b) bias profiles, (c) RMSE and (d) bias difference profiles, and statistically significant (e) RMSE and (f) bias differences. The shading in the lower four panels corresponds to the threshold criteria summarized in Table 1. All bias differences are computed using the absolute value of the average bias.

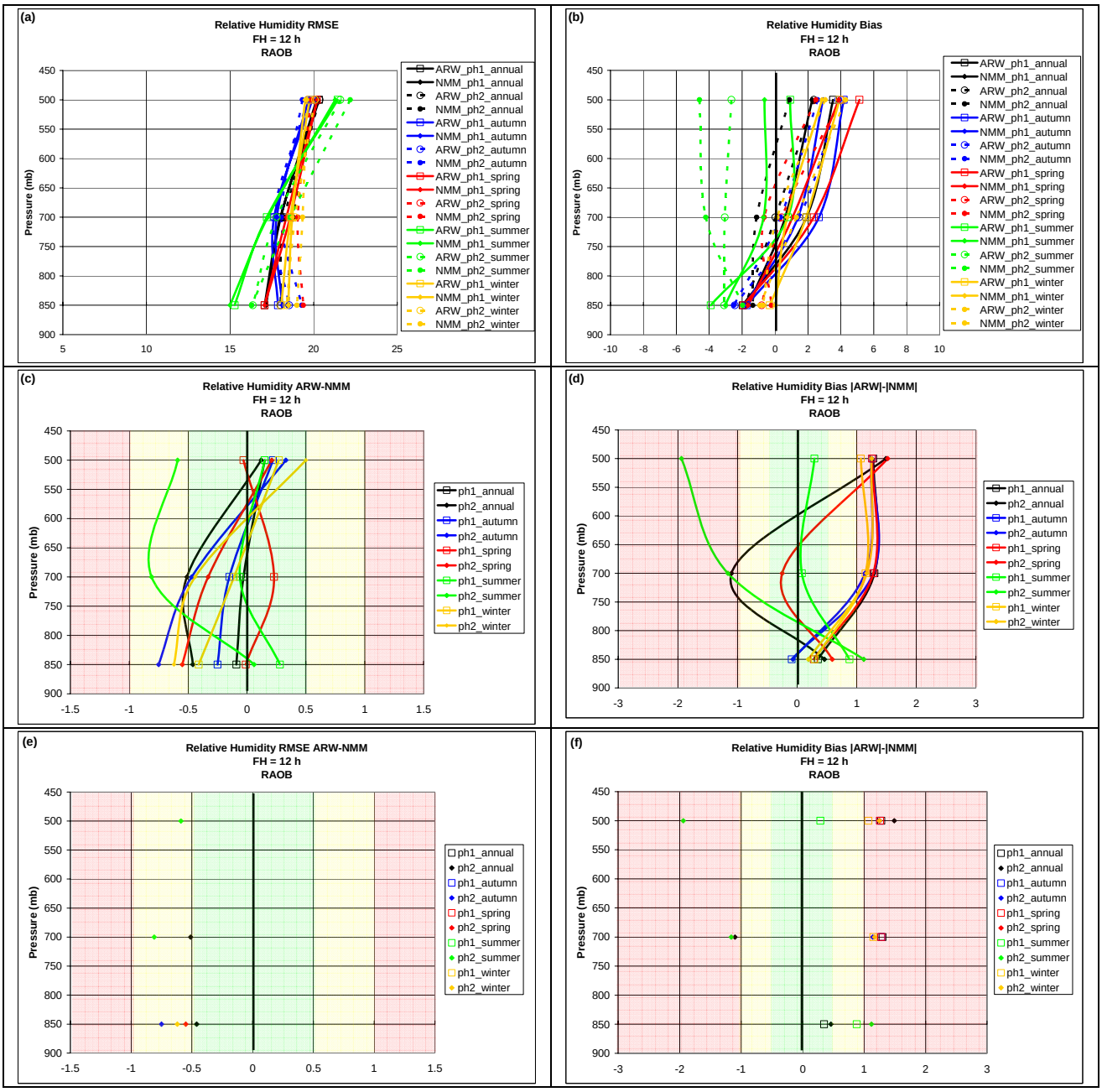

Figure 17: Same as 16 except for the relative humidity verification statistics for forecast hour 12 (FH = 12 h).

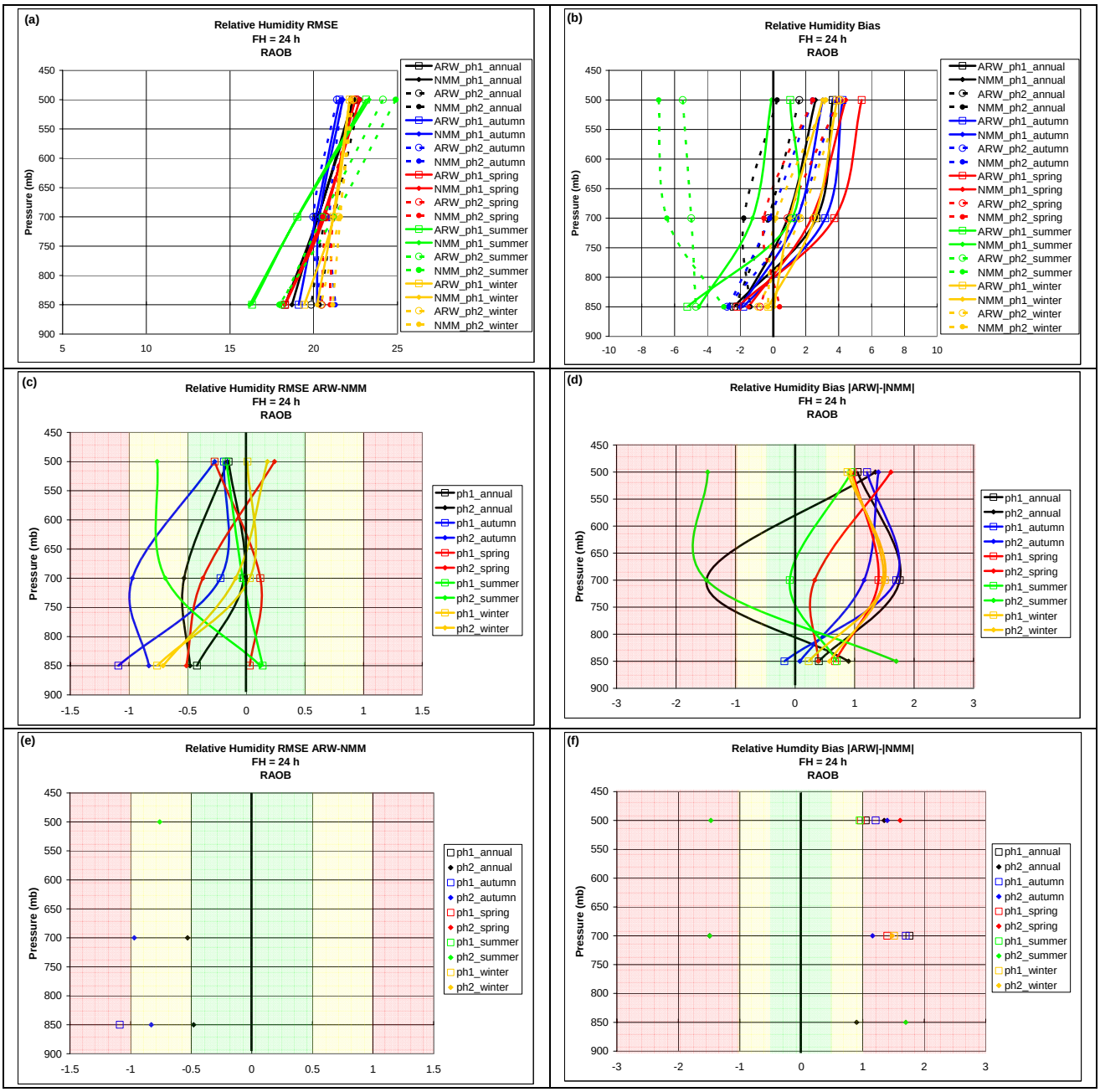

Figure 18: Same as 16 except for the relative humidity verification statistics for forecast hour 24 (FH = 24 h).

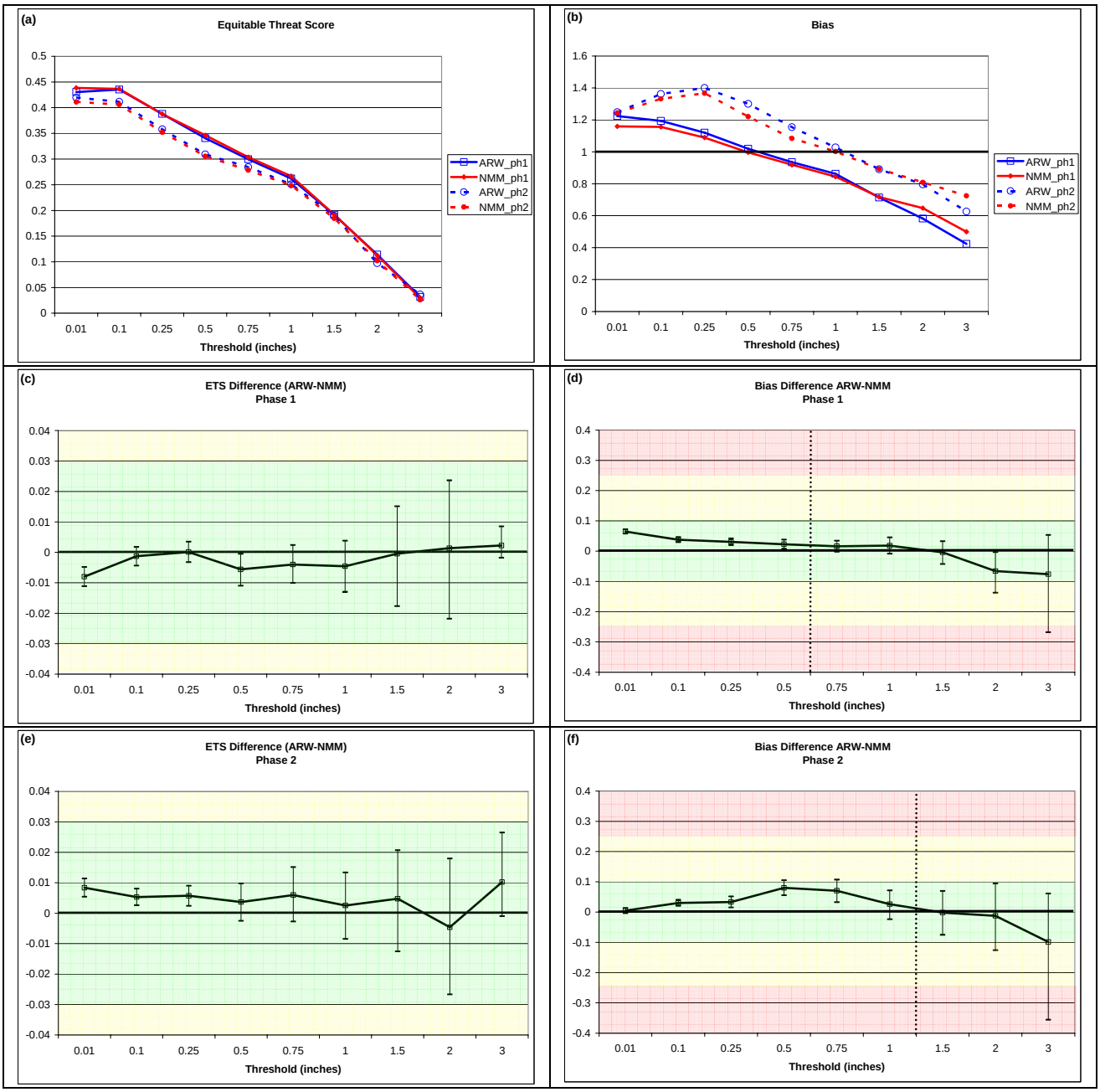

Figure 19: Precipitation verification statistics for 24-h accumulations: ETS (a), bias (b), difference between ARW and NMM ETS for phase 1 (c), bias for phase 1 (d), ETS for phase 2 (e), bias for phase 2 (f) with 95% confidence intervals. The shading in the lower four panels corresponds to the threshold criteria summarized in Table 1. The bold dotted lines in panels (d) and (f) demarcate the transition from bias greater than one to bias less than one.

Appendix A: Summary of Data Missing from WRF Core Test Archive

Forecast cycle	Missing Data	Physics Package	Reason for Missing Data
Winter			
2006012600	NMM	ph1 and ph2	forecast failed to complete
2006012812	NMM and ARW	ph1 and ph2	missing RUC input
2006012900	NMM and ARW	ph1 and ph2	missing RUC input
2006020500	ARW	ph2	forecast failed to complete
2006021500	3-h QPF Verification	ph1 and ph2	corrupt observation data
2006021512	3-h QPF Verification	ph1 and ph2	corrupt observation data
Fall			
2005111012	NMM and ARW	ph1 and ph2	missing RUC input
2005110612	ARW	ph2	forecast failed to complete
2005111600	ARW	ph2	forecast failed to complete
Summer			
2005071712	NMM and ARW	ph1 and ph2	missing RUC input
2005072300	ARW post-processed files after fhr 03	ph2	post-processor crash related to small soil moisture values
2005072612	ARW post-processed files after fhr 21	ph2	post-processor crash related to small soil moisture values
2005072700	ARW post-processed files for fhr 24	ph2	post-processor crash related to small soil moisture values
2005072800	ARW post-processed files after fhr 21	ph2	post-processor crash related to small soil moisture values
Spring			
2006032500	NMM and ARW	ph1 and ph2	corrupt RUC input
2006032912	24-h QPF verification	ph1 and ph2	missing RFC data
2006040200	NMM and ARW	ph1 and ph2	corrupt RUC input
2006041300	NMM and ARW	ph1 and ph2	missing NAM input
2006041712	NMM and ARW	ph1 and ph2	missing RUC input
2006041800	NMM and ARW	ph1 and ph2	missing RUC input
2006042300	incomplete sfcupa verification	ph1 and ph2	missing RUC prepbufr files
2006042312	incomplete sfcupa verification / 24-h QPF verification	ph1 and ph2	missing RUC prepbufr files / missing RFC data