31 August 2006 (links to all reports added 7 September 2006)

Memorandum to: **Steve Lord**
NCEP Environmental Modeling Center, Director
Nelson Seaman
NOAA/NWS/OST WRF Program Coordinator

From: **Stan Benjamin**
WRF Rapid Refresh coordinator for NOAA
NOAA/ESRL/GSD Assimilation and Modeling Branch, Chief
John Brown, NOAA/ESRL/GSD/AMB

Subject: **WRF Rapid Refresh dynamic core evaluation and recommendation**

Executive Summary, Recommendations:

- Based on the recent 9-month evaluation, we in the Rapid Refresh development group in NOAA/ESRL/GSD recommend, by a slight margin, the ARW core over the NMM core for the initial operational Rapid Refresh implementation planned for 2008.
- Some significant advantages were evident for one core or the other, dependent on variable or vertical level, with a slight edge for ARW overall, but we judged that there was no strong overall advantage for either.
- Therefore, we will fully support your decision, regardless of which core you choose. We urge your careful consideration of the comparison results below.

- This recommendation is based on the 9-month-long WRF core-test evaluation and the combined recommendations and comments primarily from these organizations:
    - NOAA/ESRL/GSD
    - Developmental Testbed Center (NCAR and NOAA/ESRL/GSD)
    - Product Development Teams for aviation weather products funded through the FAA Aviation Weather Research Program

- Our recommendation is for the *initial* operational implementation of Rapid Refresh currently planned for 2008. NOAA's purpose for use of WRF in NCEP operational models is to accelerate transition from research into operations on an ongoing basis. Yearly update packages will be made to the Rapid Refresh code, similar to the process for the current Rapid Update Cycle (RUC) model and other NCEP operational forecast models. Given the likelihood that important further developments will emerge from the WRF modeling community, we recommend that a re-evaluation of the WRF-RR dynamic core should be conducted every 2 years based on WRF community developments.

- As a result of the WRF-RR core-test evaluation, a full set of physical parameterizations from two different suites are now available for use with either the NMM or ARW cores, including the parameterizations likely required for the operational Rapid Refresh.

Outline of report

1. Introduction
2. Ground rules
3. Procedure
    a. Code modifications to WRF model necessary for evaluation
    b. Experimental design
    c. Model configuration
    d. Post-processing
    e. Verification
4. Background information for results
    a. Strongly constrained experiments
    b. Issues with different observation types for verification
    c. Significance of statistical differences
5. GSD evaluation
    a. Statistical evaluation
    b. Visual evaluation
6. Summary on core-test preferences from WRF-RR core-test reports other than GSD

    Turbulence PDT (summary, full report)
    In-flight Icing PDT (summary, full report)
    Convective Weather PDT (summary, full report)
    National Ceiling/Visibility PDT (summary, full report)
    Developmental Testbed Center (summary, full reports – Nance, Weatherhead)
    NCEP Aviation Weather Center (comments)
    NCEP Storm Prediction Center (comments)
7. Discussion and conclusions

Appendices

a. Modifications to WRF model necessary for the Core Test Pertaining to Physics Schemes
b. WRF namelists (ARW, NMM) used for Phase 2
c. Annual summary statistics – 24-h forecasts, eastern verification area
d. Summary comments from AWRP PDTs , DTC, and NCEP/AWC and NCEP/SPC on WRF-RR core-test evaluation
e. Special experiments to explore cause of ARW – NMM differences in character of small-scale features

1. Introduction

Toward a NOAA goal to accelerate transition of development from the research community into operational environmental models, NOAA/NCEP and NOAA/ESRL/GSD agreed in 2002 to incorporate a version of the Weather Research and Forecast (WRF) model into a future version of the current Rapid Update Cycle (RUC) model, to be called the *Rapid Refresh (RR)*. The WRF model has a number of options, including two primary versions of its dynamic core: the Advanced Research WRF (ARW) and Non-hydrostatic Mesoscale Model (NMM) cores.

In accord with this agreement, a 9-month objective evaluation was conducted by NOAA/ESRL's Global Systems Division in collaboration with the NOAA/NCAR Developmental Testbed Center and NCAR. This testing of the ARW and NMM dynamic cores was unique from any previous WRF testing in that well-controlled experiments were designed to isolate differences arising from the cores themselves. This was only possible with substantial changes to the WRF model itself. The *goal of this evaluation* was to *determine the best WRF dynamic core for the specific aviation and severe weather applications planned for the Rapid Refresh* at the expected initial Rapid Refresh horizontal resolution (~13km).

The effort to develop a version of the WRF model specifically suitable for the Rapid Refresh began in 2003 in NOAA. Since October 2003, almost 3 years ago, NOAA/ESRL/GSD (previously Forecast Systems Laboratory) has been testing and modifying evolving versions of the WRF model with RUC initial conditions in real-time tests at 13-km resolution.

After selection of a dynamic core for the Rapid Refresh WRF, GSD will prepare for an extended 9-12 month in-house testing period at GSD. The following schedule is contingent on FAA and NOAA funding in FY07-08.

- Full Rapid Refresh testing period – October 2006 – June 2007.
    - Use of WRF model with chosen core
    - Use of likely Rapid Refresh physics configuration
    - Cycling using GSI (new analysis – Gridpoint Statistical Interpolation) assimilation. [Per decision by GSD/FSL in June 2005 to use GSI in the RR]
    - Full Rapid Refresh domain (including Alaska, Canada, etc.)
    - Real-time testing at GSD augmented by some retrospective testing
- Transfer of RR system to NCEP for EMC testing – July-August 2007
- Pre-implementation testing at NCEP starts – September 2007
- Operational implementation – March-June 2008

2. Ground rules

A set of ground rules was established by NOAA/ESRL/GSD in its 15 December 2005 document entitled "WRF-Rapid Refresh core comparison plans – planned collaboration between ESRL/DTC/AWRP-PDT".

- In the event of very similar overall performance for aviation applications (see appendix), the NMM core will be selected. Reason: More NOAA leverage with ongoing NCEP/EMC effort to further develop the NMM. Leverage from the larger WRF community is still important for the success of the WRF model for the Rapid Refresh.
- GSD RUC/Rapid Refresh group has final responsibility for recommendation on core selection to be presented to NCEP/EMC (Stephen Lord, Director) and the NOAA/NWS/OST WRF Program Coordinator (Nelson Seaman).
- Results from other WRF core testing (not specifically for RR core) from the WRF community may be relevant and should be considered, if available.
- NOAA/NWS/NCEP/EMC has final responsibility for the decisions on the Rapid Refresh implementation, as it already has for the RUC. As with the RUC, recommendations for the Rapid Refresh configuration from GSD including the AWRP development community will weigh heavily in EMC decisions.

3. Procedure

The WRF-RR core-test project has consisted of 3 phases: preparation (November 2005 – April 2006), forecast tests (late April – late May), and evaluation (May-August 2006).

Since the goal of this WRF-RR core-test experiment was to determine the best configuration for the Rapid Refresh application, its experimental configuration was designed to be as similar as possible to that of the current Rapid Update Cycle, including ~13k-km resolution, emphasis on short-range forecasts, and use of RUC initial conditions including 3-d hydrometeor fields. The procedure was designed to conduct experiments as broad in scope as possible, given resources available, and to ensure controlled experiments, as much as possible. The procedures are broken down into model code modification and experimental design.

A. Code modifications to WRF model necessary for evaluation

The WRF-RR core-test evaluation was conducted with two sets of physical parameterizations, the NCEP physics suite currently used with the WRF-NAM (North American Mesoscale) operational model, and a different set of physics developed for aviation and severe weather forecasting applications with the RUC model (Table 1).

Parameterization	Phase 1 – NCEP physics suite	Phase 2 – RUC physics suite
Explicit clouds	Ferrier	Thompson-NCAR
Sub-grid-scale convection	Betts-Miller-Janjic	Grell-Devenyi
Land-surface	NAM/F77 version of Noah ("99" LSM)	RUC-Smirnova
Turbulence mixing	Mellor-Yamada-Janjic	Mellor-Yamada-Janjic
Radiation	Longwave/Shortwave – GFDL	Longwave /Shortwave - GFDL

Table 1. Physical parameterization suites used in the WRF-RR core-test evaluation

At the beginning of the WRF-RR core-test evaluation project (consisting of both preparation and actual test), we expected that implementation of the Phase 2 physics suite into the WRF model for both the ARW and NMM cores would be the most daunting task of the evaluation. As it turned out, even the Phase 1 physics did not fully work with the ARW core. Table 2 indicates the physics availability at the beginning of the testing phase WRF-RR core evaluation project.

Parameterization	NMM	ARW
Ferrier microphysics	✓	Radiation problem
NCAR-Thompson microphysics	-	Old version in WRFv2.1
Mellor-Yamada-Janjic PBL	✓	✓
	- but changed by EMC – Feb 2006 - but designed for use with Ferrier microphysics – modifications required for Phase 2 physics	
Betts-Miller-Janjic convection	✓	✓
Grell/Devenyi convection	-	✓ - old version
Option 99 LSM	✓	-
Noah (F90) LSM	-	✓
RUC/Smirnova LSM	-	✓
RUC initial conditions	-	✓

Table 2. Physics availability *at beginning* of WRF-RR core-test project. Checkmarks indicate availability (sometimes with limitations), and dashes indicate nonavailability.

After a considerable effort led by GSD but with significant contributions by NCEP/EMC, NCAR/RAL, and NCAR/MMM in the core-test preparation phase, the physics availability was significantly enhanced for the WRF model to that shown in Table 3. This effort is described in more detail in Appendix A.

Parameterization/capability	NMM	ARW
Ferrier microphysics	✓	✓
Thompson microphysics	✓ - New version	✓ - New version
Mellor-Yamada-Janjic PBL	✓	✓
Betts-Miller-Janjic convection	✓	✓
Grell/Devenyi convection	✓ - New version	✓ - New version
Option 99 LSM	✓	✓
Noah (F90 – option 2) LSM	-	✓
RUC/Smirnova LSM	✓	✓
RUC initial conditions	✓	✓

Table 3. Physics or initial conditions availability resulting from the WRF-RR core-test project. Bold, blue checkmarks indicate new or modified availability from the core-test project.

These modifications are all being transferred to the WRF Repository (by Grell, Peckham, Smirnova – GSD), allowing use by the WRF community.

B. Experimental design

The experimental design is described in more detail by the DTC report on the WRF-RR core-test evaluation (Nance et al). The highlights are described below:

- Conduct evaluation over four separate months, one in each season, to evaluate differences across an annual cycle:
  - Fall 1-30 Nov 2005
  - Winter 15 Jan – 15 Feb 2006
  - Spring 25 Mar – 25 Apr 2006
  - Summer 15 Jul – 15 Aug 2005
- Two runs per day, initialized at 0000 and 1200 UTC, 24h duration

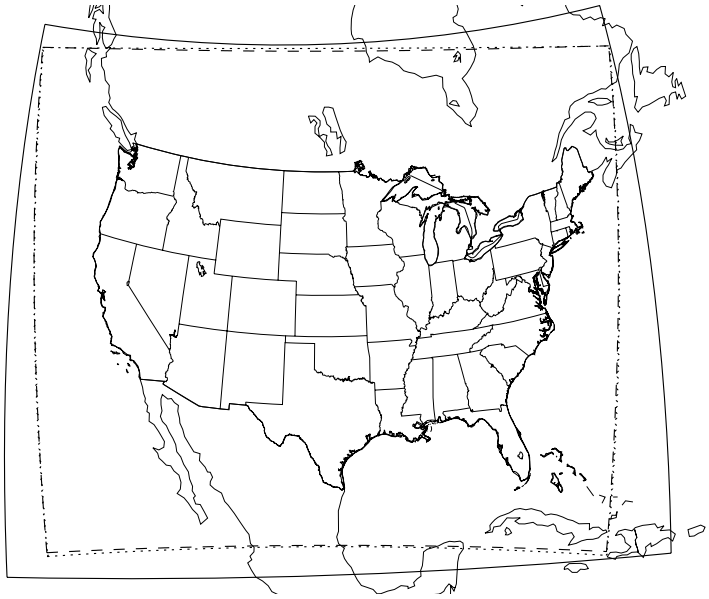


Figure 1. ARW and NMM forecast domains for WRF-RR core test. Solid line is RUC13 451x337 domain. Dashed and dotted lines are for ARW (Lambert conformal) and NMM (rotated lat/lon) domains. (From DTC report – Nance et al.)

- Horizontal domain

Nearly matching sub-domains of RUC CONUS domain (Fig. 1, more detail in DTC report -- Nance et al.) were developed by DTC for the WRF-RR core test experiments. Some small differences were necessitated by the different map projections available for the ARW and NMM cores.

- Initial conditions for atmosphere and land-surface, including SST –

RUC13 native coordinate data from GSD dev13 RUC runs, which incorporates the NCEP SST-14km daily updated fields. Since RUC analyses include 3-d multi-hydrometeor cloud fields with assimilation of GOES cloud-top and METAR (surface) cloud observations, it was considered essential to use RUC initial conditions, including 3-d cloud fields. The WRF SI (see below) can currently provide initial 3-d cloud fields (avoiding cloud/precipitation spin-up) only with RUC grids.

- WRF Standard Initialization (SI) modifications

Use SI (including new necessary changes) to properly interpolate RUC initial conditions, including hydrometeor fields and land-surface fields (soil temperature and moisture, snow water equivalent, snow temperature, ice cover) into both NMM and ARW.

- Lateral boundary conditions –
    - NAM AWIPS 212 grid (40km) using NAM run initialized 6h previous to current WRF initial time (matching operational RUC BC).

Model configuration

- Model physics for Phase 1 and 2 – As shown in Table 1 above.
- Horizontal resolution – approximately 13km. More detail in DTC report.
- Time-step.

- ARW – 72 s for dynamics, 72 s for physical parameterizations other than radiation.
- NMM – 30 s for dynamics, 60 s (2 dynamic time steps) for physical parameterizations other than radiation.
- These time step lengths were shown to be the maximum stable values in testing for the other configurations

- Vertical levels - 50, distribution of levels set similarly for ARW and NMM.

(Cannot be equal, since ARW uses a sigma configuration, and NMM uses a hybrid sigma-pressure vertical coordinate, with the switch to isobaric levels at ~420 mb.)

- Top of model domain - 50 mb

(will be changed to higher level for subsequent Rapid Refresh testing with GSI assimilation)

- Upper boundary condition -
  - NMM – default (with divergence damping)
  - ARW – upper-level damping layer (damping coefficient set at 0.02 after experimentation, damping layer depth = top 5 km of model domain)

[Note: The WRF namelists used for ARW and NMM for Phase 2 physics are provided in Appendices B1 and B2. Namelists for Phase 1 physics differed only in the options for physical parameterizations.]

- Terrain elevation specification

Due to differences in the SI software for vertically interpolating RUC native-coordinate data to the NAM and ARW vertical coordinates, some differences resulted in terrain elevation between the NMM and ARW. NMM terrain was limited to values similar to those in the RUC terrain. This was due to SI software design, and results in NMM terrain somewhat smoother than terrain using default smoothing settings. To mimic the smoothness in the NMM terrain when using RUC initial conditions through SI, a higher smoothness value was used with the ARW terrain (TOPTWVL_PARM_WRF = 4). Nevertheless, some differences remained, as shown below in Fig. 2. Therefore, we have generally preferred verification results from the Eastern US verification domain vs. the Western US domain or the full ("National") verification domain.

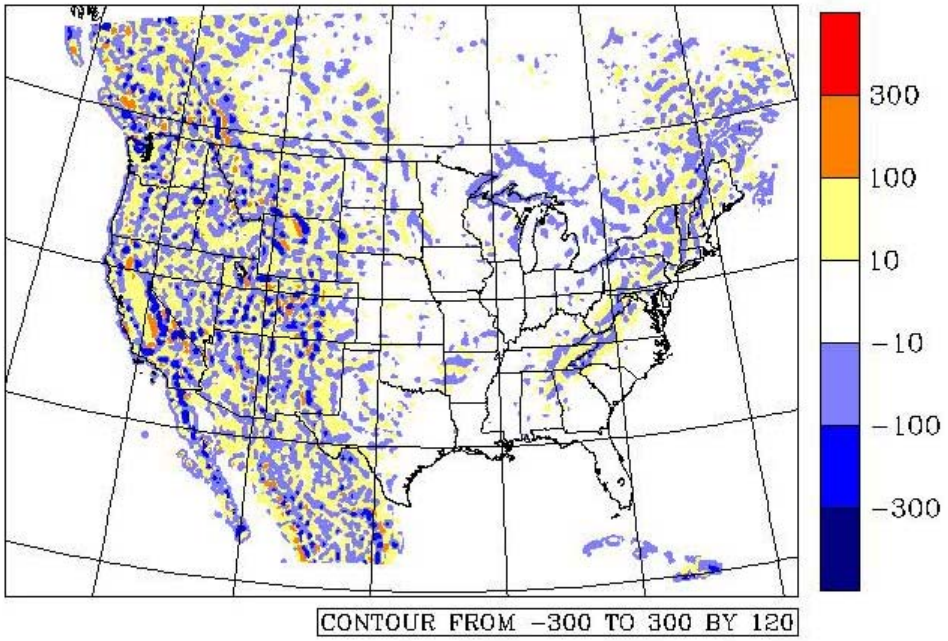


Figure 2. Terrain difference – ARW value minus NMM value. (Nance)

Post-processing

The WRF post-processing routine (*WRFpost*, largely developed at NCEP) was used to produce isobaric grids and other diagnostic products on the mass-coordinates of the native grids for ARW and NMM models (see DTC report for more detail). Then, the NCEP program *copygb,* was used to interpolate NMM and ARW output to the RUC grid (451 x 337 grid points), flagging grid points with missing values as needed.

E. Verification

Verification of WRF-RR core-test forecasts was performed against 5 kinds of observations, all through the NCEP-WRF verification package: rawinsonde, aircraft, profiler, surface, precipitation (3-h, 24-h). Verification statistics (based on observation-forecast differences) were calculated for the full model domain but also over two regional sub-areas (Fig. 3). Additional detail on this is provided in the DTC report on the WRF-RR core-test evaluation.

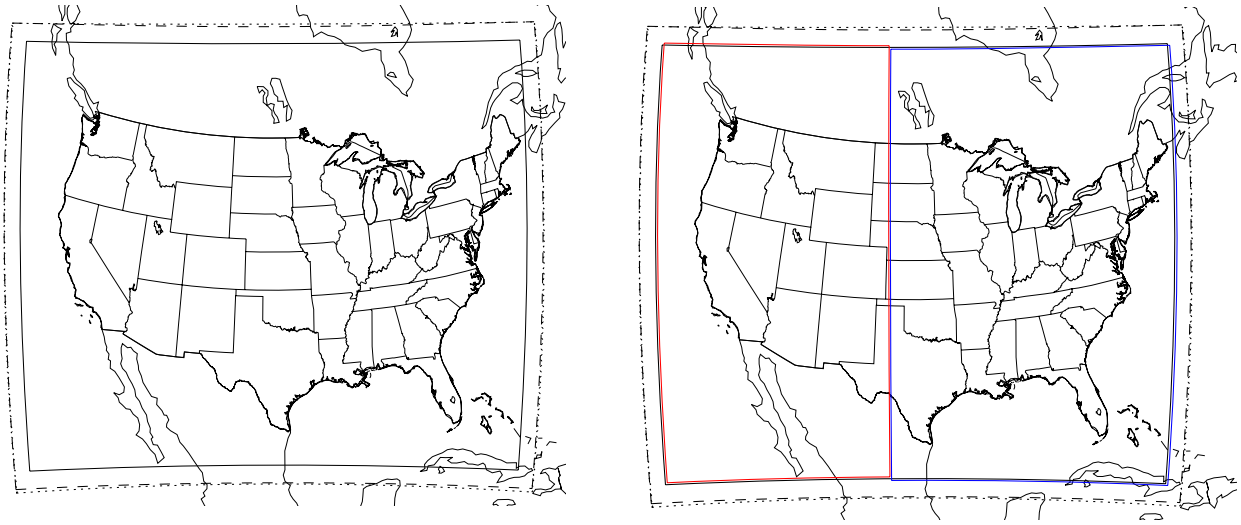


Figure 3. Verification areas, over a) full CONUS area (just inside full model domain) (on left) and b) western and eastern regional verification areas (on right). (Nance)

GSD also ran the RUC verification package used for previous RUC pre-implementation testing to ensure that it gave approximately the same results as the NCEP-WRF package.

4. Background information for results

We start this discussion of results in the WRF-RR core-test comparison by explaining important background information to allow the reader to understand the context.

A. Strongly constrained experiments:

The WRF-RR core-test study strongly constrained experimental design, more so than other model comparison studies done in the past at NCEP. The constraints:

- No cycling, same initial conditions for each run
- Forecasts only out to 24h
- Same lateral boundary conditions from NAM for each pair of experiments, and relatively "small" CONUS sub-RUC-size domain, further constraining core-test differences
- Differences must be calibrated consistent with the constrained design.

Seemingly "small" differences without cycling in the sub-RUC-size domain would amplify with cycling, allowing accumulation of differences over a month-long test period. Core-related differences would further amplify if comparison experiments were conducted over the full North American domain planned for the Rapid Refresh. The subjectively significant criteria shown below were estimated with these constraints in mind.

B. **Issues with different observation types for verification**

- Rawinsonde

Rawinsonde wind observations have larger observational errors than automated aircraft winds, especially above the mid-troposphere. Moreover, rawinsonde wind observations are often missing in the upper troposphere and lower stratosphere in strong wind situations due to "blow-off" resulting in too low an elevation angle for balloon locations to be accurately traced for non-GPS rawinsondes. Aircraft wind observations are not susceptible to this problem and are equally available and accurate in conditions of any wind speed. No rawinsonde verification is available in this data set below the 850 hPa level, whereas aircraft observations above the surface are available in the 1000-850 layer. Conversely, due to the design of this verification (originally from NCEP), no aircraft data are available above 200 hPa (~FL385) whereas rawinsonde observations are available up to 150 hPa (~FL450).

- Aircraft

Aircraft observations are not evenly distributed in time or space. Due to the commercial airline flight structure based on public needs, there are many more automated aircraft reports available over North America at 0000 UTC than at 1200 UTC. Therefore, aircraft-based verification statistics over 24-h periods are skewed toward daytime and early evening hours. Similarly, aircraft-based statistics over the CONUS area are skewed toward major hubs in the eastern and central United States.

- Profiler

Profiler observations were not used due to an error in the NCEP observation processing wherby height-pressure mapping used the Standard Atmosphere rather than the local height-pressure profiles.

C. ***Significance of statistical differences***

Obviously, the differences between forecast skill for the ARW vs. NMM experiments must be judged for their significance. GSD and other core-test participants agreed that criteria and procedures should be established *a priori* for ***two types of significance tests****.* The first is the classic, formal "***statistical significance***". The second is based on magnitude of statistical differences that would be clearly evident in visual comparison of horizontal plots of forecasts from the two models, and we call this "***subjective significance***". These terms are defined in sections 4.C.1 and 4.C.2 below. After defining these terms, we then introduce a measure in section 4.C.3 to identify **"significant seasonal differences" (SSDs)** for a given variable at a given level showing both formal statistical and subjective significance.

1) Statistical significance

"Formal" statistical significance evaluation of WRF-RR core-test results was calculated by Betsy Weatherhead and Greg Noonan (NOAA/ESRL/GMD). Their techniques are described in detail in the DTC report. Barry Schwartz of GSD also performed an initial statistical significance comparison using the Student's t-

test score. Only Weatherhead/Noonan results are shown in this report, and a summary of their procedure is briefly described below:

- Pair-wise comparison, distributions assumed to be Gaussian and auto-regressive 1st order
- Will give different results than comparisons of means and distributions.
- Auto-correlation of differences is taken into effect.
- Separate statistics are produced for
    - 3 verification areas: CONUS, western, and eastern are examined.
    - Phase 1 and 2 physical parameterization suites
    - All four seasons are considered separately and jointly.

2) Subjective significance

As part of this project, in April 2006, GSD established subjective evaluation criteria for **magnitude of error difference** averaged over 1-month periods based on previous implementations, especially for the RUC. These criteria were based on statistical differences that would be **clearly evident in visual comparison of horizontal plots of forecasts from the two models**. Horizontal plots of forecast-minus-analysis error fields (see Fig. 11, Benjamin et al. 2004, for a specific example for 250-hPa winds) show that errors are only large in certain regions, usually near significant weather events. **RMS errors include areas with very accurate forecasts and very small error, but a few areas with much larger error. The criteria below are set to detect local areas of much larger error differences**.

These subjective significance criteria were established by variable types as follows:

🟢 Difference considered as insignificant

🟡 Difference is of concern ("yellow zone")

🔴 Difference is of *serious* concern ("red zone")

These criteria were to be applied with the following conditions:

- Differences should be consistent over the majority of individual verification times over a month-long seasonal period.
- Differences should be retained as important only if they are also found to be statistically significant (see next section).

Subjective evaluation criteria – upper-air	**Wind** - any level (850-150 hPa) – RMS *vector* error	**Temperature -** any level (850-150 hPa) – RMS error	**RH** - 850-500 hPa – RMS error
●	< 0.10 m/s	< 0.1 K	< 0.5 %
●	0.10-0.25 m/s	0.1-0.2 K	0.5-1.0 %
●	> 0.25 m/s	> 0.2 K	> 1.0 %

Table 4. Subjective significance criteria for upper-air verification for wind, temperature and RH.

RH forecast verification was considered suspect above 500 hPa due to lower rawinsonde accuracy at temperatures < -25 ^{o}C. Verification over the Western verification area was considered suspect at 850 hPa due to proximity to surface and effects of extrapolation below ground.

No subjective significance criteria were set for *bias* for any of these variables (wind, temperature, RH). The effects of bias are included in the RMS error differences, since RMS difference was used instead of standard deviation (s.d.) difference, which eliminates bias. We consider biases in the GSD statistical evaluation below in section 5.A, but less heavily than RMS error.

In the GSD analysis of **upper-air verification (section 5),** vertical lines are plotted (e.g., Fig. 5) corresponding to the **"yellow zone"** and **"red zone"** subjective significant levels**.**

Subjective evaluation criteria – surface	**Wind** - RMS *vector* error	**Temperature -** – RMS error	**RH** - RMS error
●	Not established	< 0.2 K	Not established
●		0.2-0.5 K	
●		> 0.5 K	

Table 5. Subjective significance criteria for surface verification.

We in GSD considered the surface verification to be somewhat less important than the upper-air verification, due to its susceptibility to error in reduction from the lowest sigma level to 2-m AGL temperature and 10-m wind level, unless it was consistent with 850-hPa rawinsonde or 1000-850-hPa aircraft verification.

Adjustments to the reduction procedure are relatively easy to implement at a later stage in Rapid Refresh development.

Subjective evaluation criteria – 24h precipitation	**Precipitation – equitable threat score** up to 1.0" / 24h	**Precipitation – bias –** up to 1.0" / 24h
●	< 0.03	< 0.1
●	0.03 – 0.05	0.1-0.25
●	> 0.05	> 0.25

Table 6. Subjective significance criteria for precipitation verification.

3) Significant seasonal differences

To isolate the most meaningful differences overall for the WRF-RR core recommendation, we decided to compile *month-long differences for a given level and variable that meet criteria in both subjective significance and statistical significance* to be considered as "**significant seasonal differences**", or **SSDs**. SSDs, meeting both "subjective" and "statistical" criteria as defined in 4.C.1-2 *for any season,* are considered important since they would also likely be noticed by forecasters or aviation forecast users in those seasons. *Annual* averages (Appendix C) almost always show smaller differences, all less than *subjectively* significant criteria for concern (section 4.C.2 – "yellow zone") than seasonal differences.

SSDs were tabulated for each variable/level and for both Phase 1 and 2 physics in this report using **24-h forecasts** (generally stronger signal than 12-h forecasts) in the **eastern verification area** (also stronger signal in general, attributed to greater distance from the western boundary where common lateral boundary conditions more strongly constrain model differences (see section 4.A). Results from the western verification area are also problematic in that its 850-hPa temperature and RH statistics should be discounted (the 850-hPa surface often lies below ground) and due to the more significant terrain differences in this region (Fig. 2).

5. GSD evaluation

A. Statistical evaluation

[Note: All statistical plots in this section are extracted or adapted from the DTC report section from Weatherhead and Noonan]

Overall guidelines

- We consider mandatory isobaric levels as approximately equal in importance for this test since aviation (one of the two key RUC/RR user groups) operates on height-based flight levels. The smaller mandatory pressure intervals in the upper troposphere (50 hPa) are approximately equal in height separation to those of the mandatory pressure intervals in the lower troposphere.
- Results for rawinsonde and aircraft verification should corroborate each other, subject to the error and distribution differences described in section 4.B.

- **"Significant seasonal differences" (SSDs)** were defined in section 4.C.3 as month-long statistical differences between ARW and NMM cores for a given variable and level and physics suite considered significant from both formal statistical and subjective perspectives. Again, **SSDs were tabulated for each variable** using **24-h forecasts** (stronger signal, in general) in the **eastern verification area**, as also described in section 4.C.3.

Formal statistical significance at the 95% level is identifiable in subsequent figures where horizontal bars identifying 2-sigma (two standard deviations) do not touch the zero line, meaning that the model forecasts were from two different populations with a 95% certainty. **As it turned out, all month-long differences meeting subjective criteria also met formal statistical criteria, and were therefore identified as SSDs.** The opposite was also true: almost all model difference in RMS errors showing formal statistical significance also met subjective significance criteria ("yellow zone" – concern).

1) Wind

We discuss separately forecast verification for winds aloft and for winds at surface.

Wind above surface

Forecast error is calculated as a difference between forecasts and observations valid at the time of the forecast. Wind RMS vector error is shown in Fig. 4 over the full annual test period in this experiment for 4 different versions of the WRF model. RMS wind vector errors typically peak near the tropopause level, near 200-250 hPa on the average, as shown in this example. *Differences* between forecast errors using different models are shown in Fig. 4a to determine which model was more accurate, on the average. For instance, the ARW model with

Phase 1 physics (marked with a blue **x**) can be compared with the corresponding NMM model also with Phase 1 physics (marked with a red □). The difference at 200 hPa, in this case, shows that the ARW-Ph1 version was more accurate than the NMM-Ph1, according to rawinsonde observations.

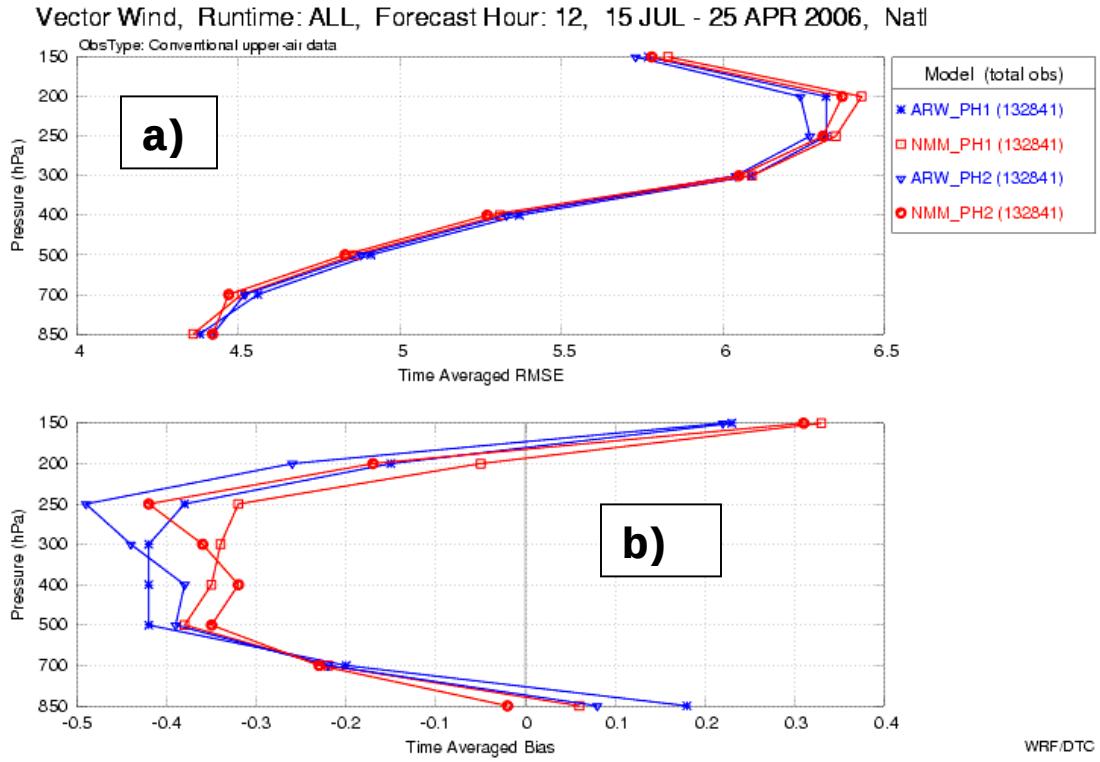


Figure 4. 12-h RMS vector wind error (a - top) and bias (b- bottom) vs. rawinsonde observations averaged over all 4 seasons for 4 different WRF model versions, ARW and NMM cores, each with Phase 1 and Phase 2 physics. Units in m/s. (From WRF-DTC verification display at http://ruc.noaa.gov/wrf/RR/testing/NCEP_verif .)

In the rest of this section, we illustrate our discussion with verification plots of these same ***error differences*** based on the analysis of Weatherhead and Noonan, presented in the DTC report. For RMS error difference plots for winds, vertical lines at ±0.1 m/s and ±0.25 m/s mark the boundaries of the "yellow zone" and the "red zone" respectively (section 4.C.2) indicating significant differences in errors between different model versions.

In Fig. 5, wind verification results show more accurate 12-h forecasts from the ARW at upper levels and, to a lesser extent, from the NMM in the lower troposphere, averaged over the 4-season test periods. These differences are muted when verified against aircraft reports: smaller advantage for the ARW at upper levels, and almost no advantage for the NMM in the lower troposphere. For 24-h forecasts (Fig. 6), the same general patterns are evident, with slightly greater difference in upper-level wind for ARW in the eastern area, and little difference in the lower troposphere (400-850 hPa). To better focus on 24-h

forecasts in the eastern verification area (as we shall do in subsequent figures in this section), we reproduce Fig. 6 in Fig. 7 showing only the eastern area. The justification for using the Eastern verification area (section 4.C.3) is evident in Fig. 6 where upper-air (200-300 hPa) differences are clearly muted in the Western verification area (shown by **W/w**), a result of constraint by the nearby inflow western boundary condition.

We next present seasonal variations for the same statistics for both rawinsonde verification (Fig. 8) and aircraft verification (Fig. 9). The NMM performance compared to that of the ARW is strongest in the summer and spring. The ARW has strongest seasons in winter and fall (November is late fall).

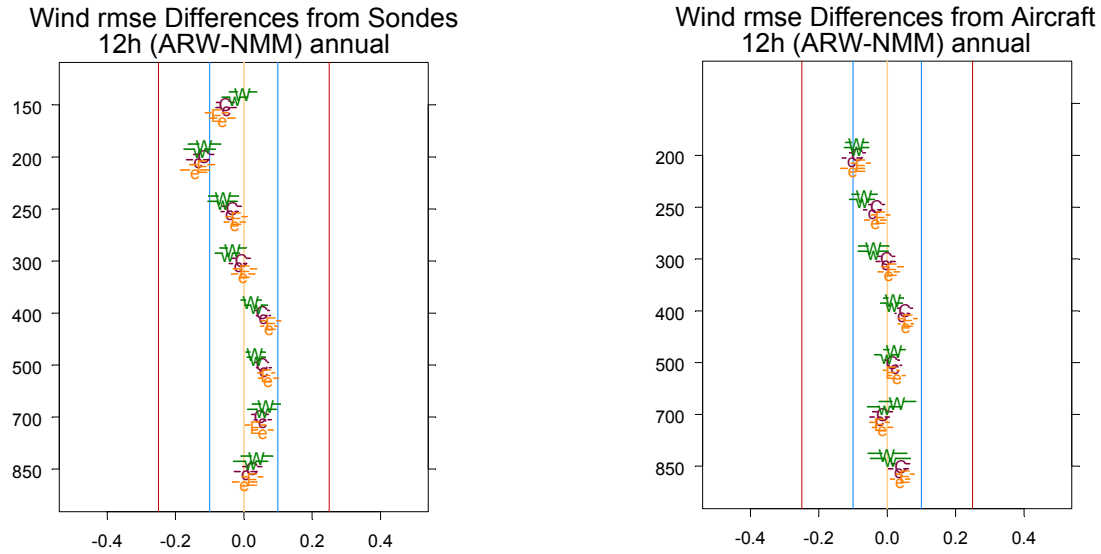


Figure 5. Wind RMS vector error ARW-NMM differences (units – m/s) over all 4 seasons for verification against observations from a) rawinsonde (left) and b) aircraft (right). The vertical axis is pressure in hPa, and the aircraft values represent layers, e.g., 200 hPa represents the 200-250 hPa level, and so on. Results are shown for 3 verification areas, C/c – CONUS, W/w – western, E/e – eastern. Upper-case letters are for experiments with Phase 1 physics, and lower-case letters are for same with Phase 2 physics. The width of the bar through each letter is for the 2-σ deviation, indicating 95% statistical significance that the forecasts are different if the bar does not intersect the zero axis. Blue and red lines are shown at "concern" (±0.1 m/s) and "serious concern" (±0.25 m/s) differences, as described in section 4.C.2.

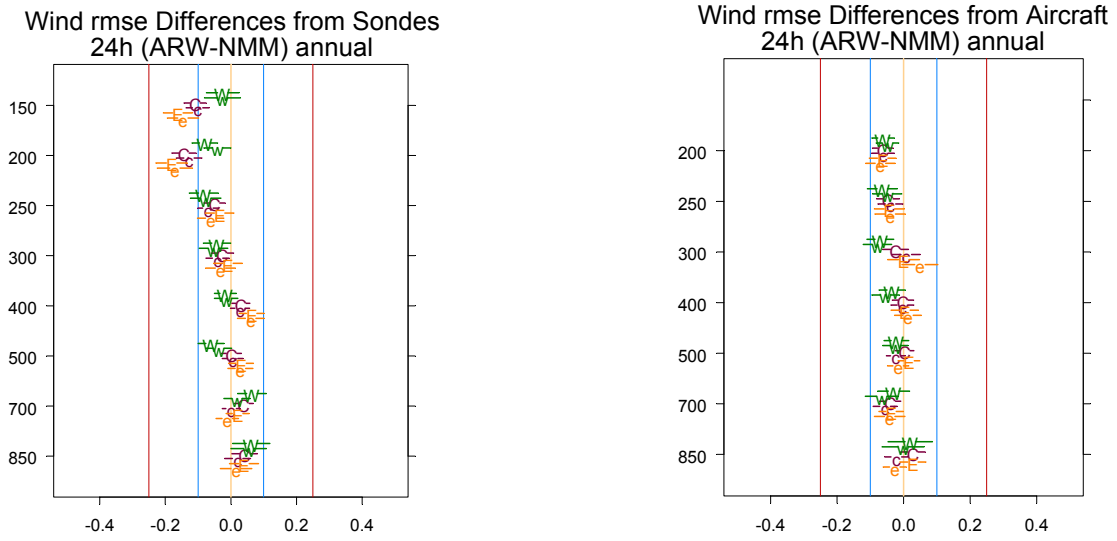


Figure 6. Same as Fig. 5 but for 24-h forecasts.

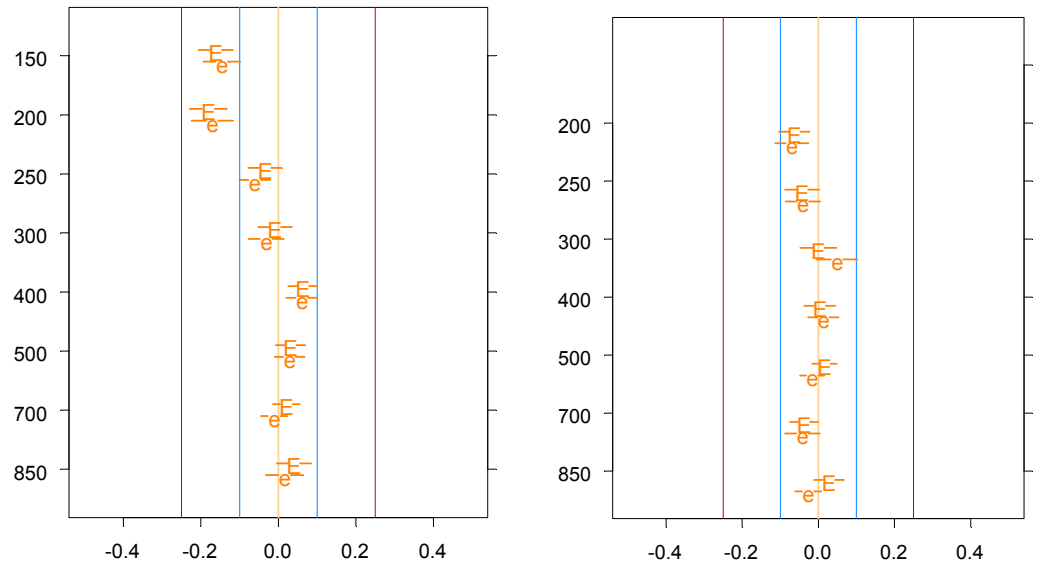


Figure 7. Same as Fig. 6 (ARW-NMM difference in RMS wind vector for 24-h forecasts, rawinsonde on left, aircraft of right), but now showing *eastern verification area statistics* only. Large case 'E' is for Phase 1 physics, and small case 'e' is for Phase 2 physics.

To summarize these results for periods in which one core or another might be considered to be noticeably better for aviation, ***significant seasonal differences (SSDs)*** (as defined in section 4.C.3) were determined. SSDs are **important error differences** over the eastern verification area for either physics suite for a given variable, level, and month-long season that met subjective evaluation criteria (section 4.C.2) and formal statistical significance (section 4.C.1). Considering two physics suites, 4 seasons, and 8 mandatory pressure verification levels, there are 64 possible SSDs for each variable.

For *rawinsonde* verification over each of 4 seasons (Fig. 8), there were 13 SSDs with lower ARW error (all at 150 hPa or 200 hPa, none in summer) and 3 SSDs with lower NMM error (2 at 400 hPa, 1 at 850 hPa) out of the possible 64. One of these events (200 hPa – winter for Phase 2 physics) exceeds even the "red zone" difference of 0.25 m/s.

For aircraft verification, there are only 7 layers (no observations at 150 hPa), leading to 56 possible SSDs per variable. When the same evaluation is completed with aircraft observations (Fig. 9), there were only 5 SSDs with lower ARW error (4 in the 200-300 hPa layers, 1 in the 850-700 hPa layer (in winter), and 2 SSDs with lower NMM error (5-2 – ARW vs. NMM in SSDs). Moreover, the "red zone" SSD noted with rawinsonde verification is reduced to a "yellow zone" with aircraft. We consider **aircraft observations to be more reliable for upper-level wind verification**, in this case, due to better representation of high windspeed situations with aircraft data (see discussion on this in section 4.B). As will be shown in the next section, smaller wind speed bias for the NMM at upper-levels is consistent with this hypothesis. Nevertheless, even with aircraft

verification, the ARW has lower wind vector errors overall, and has more favorable "significant difference events" for this variable by a 5-2 margin.

Overall, using either aircraft or rawinsonde observations, we judge the overall RMS ***wind vector*** error verification to give an **advantage to the ARW core over the NMM core.**

On the other hand, a comparison of **wind speed bias** shows that the NMM has generally a lower value than the ARW model (Fig. 10), as evident in verification against rawinsonde and aircraft observations. This was most prominent near 200 hPa and near 850 hPa. No subjective significance criterion (section 4.C.2) was established for wind speed, so we cannot estimate a score of significant difference events for wind speed bias. Moreover, wind speed bias error is included in RMS wind vector error. We still note that the **NMM core** has an **advantage** in **wind speed bias over the ARW**.. This difference was evident over all seasons (see App. C, Fig. C-4, also see DTC results from Weatherhead/Noonan).

Surface wind

Surface winds showed virtually no difference for RMS vector wind errors between ARW and NMM core experiments, but both showed a high wind bias, and this overforecasting of 10-m wind speed was more pronounced with the ARW model. This result is consistent with the 1000-850 hPa wind bias against aircraft obs shown in Fig. 8. We did not establish subjective significance criteria for surface winds due to a strong dependency on reduction to the 10-m level. However, given consistency with the bias with 1000-850 hPa aircraft wind bias, we consider this bias reflects a similar bias at the lowest few computational levels of each core.

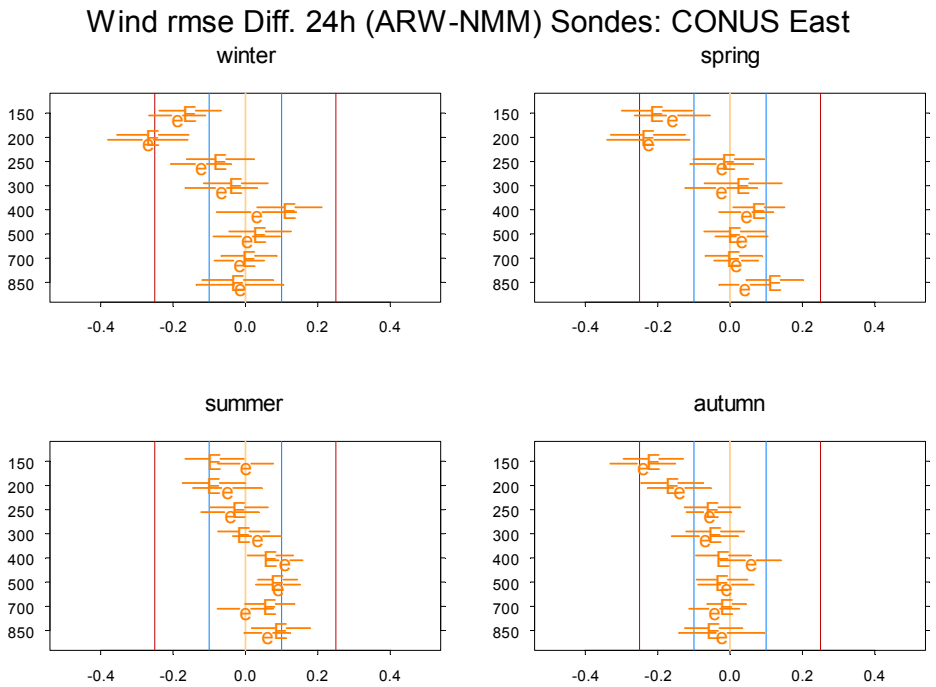


Figure 8. Wind RMS vector error ARW-NMM differences (units in m/s) verified with rawinsonde observations for 24h forecasts (eastern verification region only), but now broken down by each of the 4 month-long season periods. Blue and red lines are shown at "concern" (± 0.1 m/s) and "serious concern" (± 0.25 m/s) differences, as described in section 4.C.2. Large case 'E' is for Phase 1 physics, small-case 'e' is for Phase 2 physics.

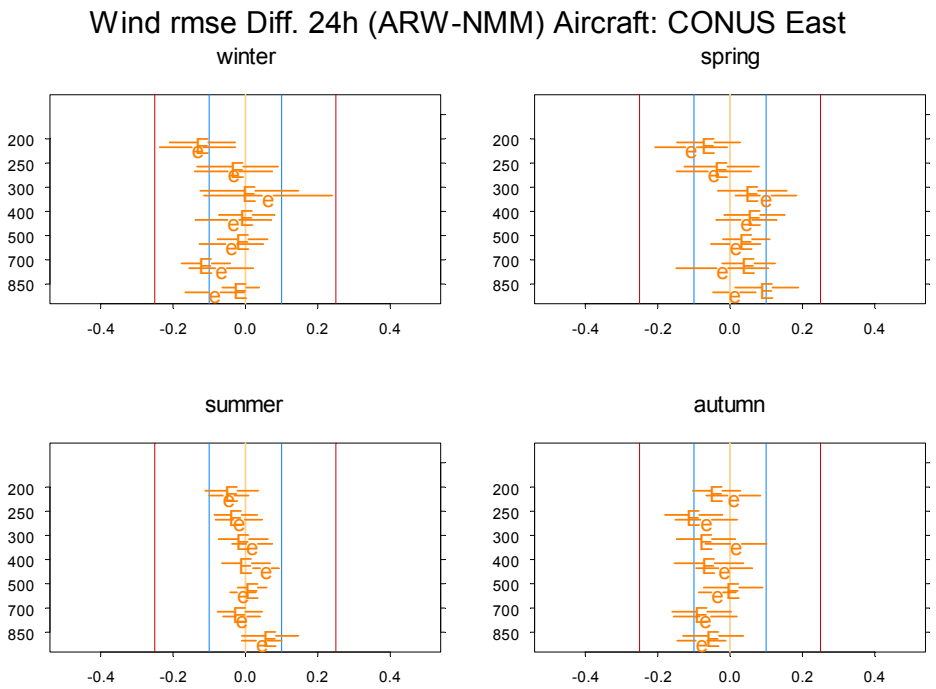


Figure 9. Same as Fig. 8 but verified with aircraft observations

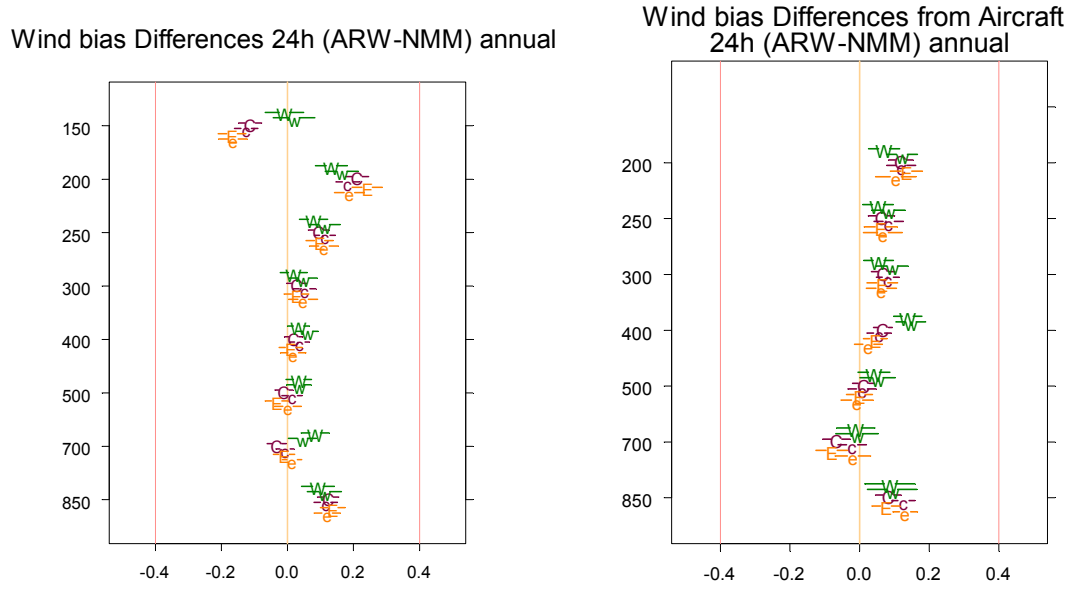


Figure 10. Wind speed bias difference for verification against a) rawinsonde (left) and b) aircraft (right). Absolute values are applied first to both ARW and NMM wind speed bias at each verification time. Positive number means that ARW absolute wind bias is larger than that from the NMM core.

2) Temperature

Temperature above surface

Temperature 24-h forecast errors are shown in Fig. 11 with rawinsonde observations and Fig. 12 with aircraft observations. Again, we focus on 24-h forecasts for the Eastern verification area to maximize model differences. Temperature forecasts were similar overall, but with some tendency for lower ARW errors for 1000-850 hPa temperatures (against aircraft data) for 3 of the 4 seasons. Using the "significant seasonal difference" (SSD) score, combining both subjective and statistical significance criteria (section 4.C.3), the rawinsonde verification in the Eastern verification area favored the NMM core by a 3-1 score (two SSDs at 150 hPa in spring for both Phase 1 and 2 out of a possible 64 SSDs). The lower 150 hPa temperature error for NMM in spring is attributable to a smaller cold bias at that level. The strongest NMM performance otherwise is near 700 hPa, but this is only for Phase 1 and Phase 2 forecasts are clearly more accurate for 700-hPa temperature.

Using aircraft verification (with no observations above 200 hPa but many in the 1000-850 hPa layer), the ARW is favored in 3 significant seasonal differences, all in the "yellow zone" and all in the 1000-850 hPa layer over 3 different seasons. For temperature, no SSDs favoring the NMM occurred (ARW – 3-0).

We judge the temperature RMS forecasts as **similar overall** but **favoring the ARW core in the lower-troposphere 1000-850 hPa layer.** Annual average temperature error differences (Fig. C.2 in App. C) also show lower ARW errors in this same layer.

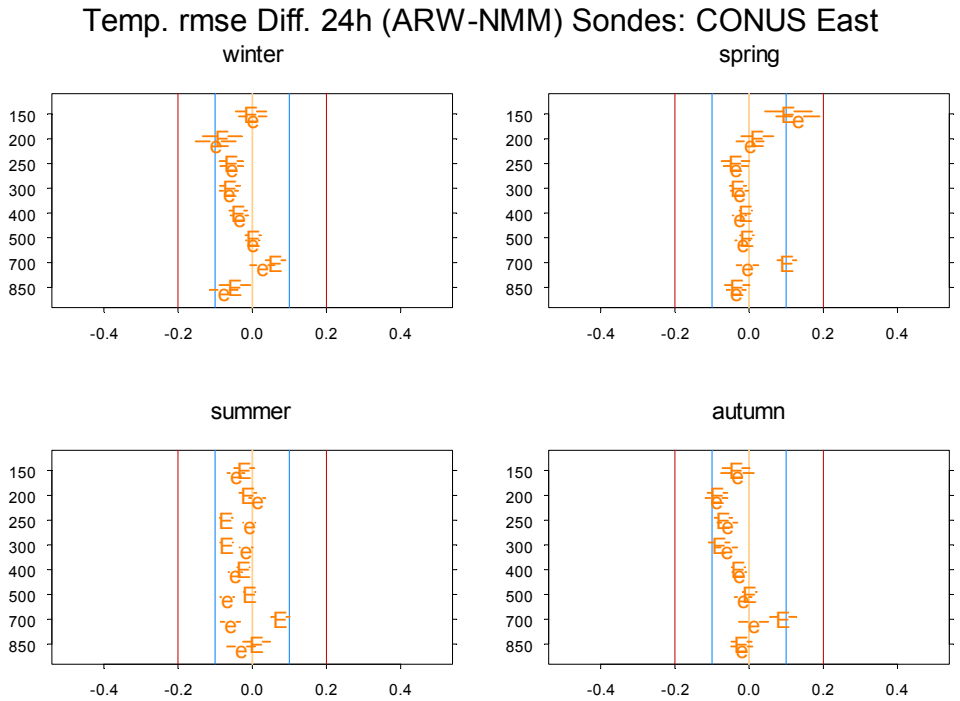


Figure 11. Temperature RMS error ARW-NMM differences (units – deg C) verified with rawinsonde observations for 24-h forecasts, broken down by each of the 4 month-long season periods. Blue lines are shown at ± 0.1 °C, corresponding to the "yellow zone" subjective significance criteria described in section 4.C.2. Red lines also shown at ±0.2 °C, corresponding to "red zone" criteria. Only eastern verification area statistics.

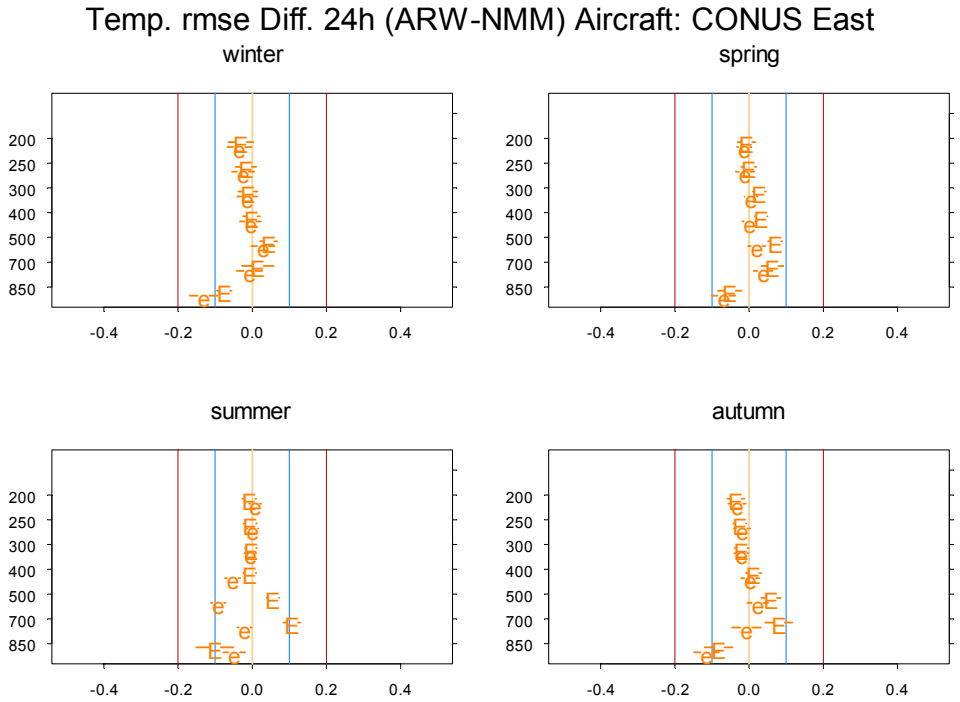


Figure 12. Same as Fig. 11, but verified with aircraft observations.

Temperature at surface

No conclusions were made regarding surface temperature verification. Differences in surface temperature forecasts were somewhat more prominent between physics suite differences than between the dynamic cores (not shown). Some small advantage was evident for the NMM in this field, somewhat less for Phase 2 physics than for Phase 1 physics. As stated in section 4.C.2, this field is dependent on reduction to 2-m level and we consider it suspect for meaningful verification.

3) Relative humidity

As shown in Fig. 13, for relative humidity aloft (above surface), the ARW core model gave lower error in colder seasons (winter, fall) at 850 hPa, whereas the NMM core gave lower errors in summer. Only 3 mandatory levels were used for RH (see section 4.C.2), and only rawinsonde observations were available (no aircraft RH observations), yielding 24 total possible SSDs. Using the SSD score, there were more SSDs favoring the ARW than the NMM (6-1). Half of the ARW-favoring SSDs occurred in the fall season, including two significant seasonal differences above the "red zone" threshold (1.0% RH).

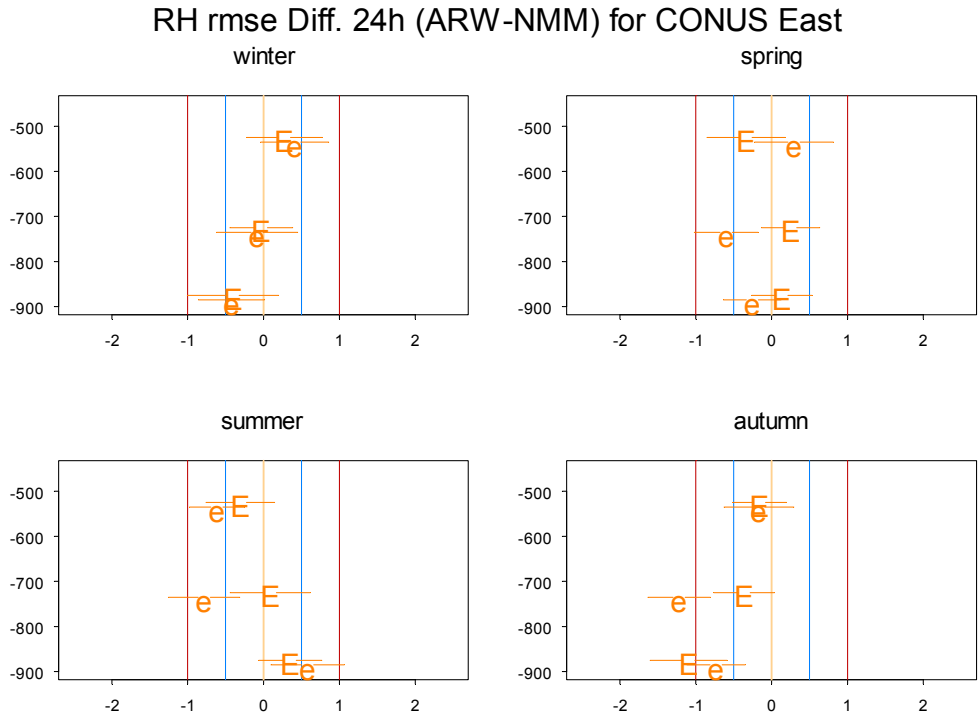


Figure 13. Relative humidity RMS error ARW-NMM differences (units – %RH, 0-100) verified with rawinsonde observations for 24-h forecasts, broken down by each of the 4 month-long season periods. Blue lines are shown at ± 0.5 %RH, corresponding to the "yellow zone" subjective significance criteria described in section 4.C.2. Red lines also shown at ±1.0 %RH, corresponding to "red zone" criteria of serious concern.

Overall, we judge the **ARW** as having **superior RH forecasts** only at 850 hPa, but this is an important level for aviation where it is linked to icing and ceiling forecasts. Annual average RH differences (Fig. C.3) show lower RH error in general for ARW forecasts.

RH bias against rawinsonde observations was also investigated. On the average, the RH bias was close to zero, except at 850 hPa where the NMM model has lower RH bias in warm seasons (about 1%RH). No conclusions were drawn from RH at the surface since it is dominated by temperature bias at the same level.

4) Precipitation

A summary of the total 4-season precipitation forecast skill (equitable threat score (ETS) and bias) is shown in Fig. 14. The results show that most of the differences are due to different suites of physical parameterizations and not to the dynamic cores. The **precipitation bias** (1.0 – no bias) is **slightly better for the NMM** for both Phase 1 and 2 physics although not at the "yellow zone" concern level.

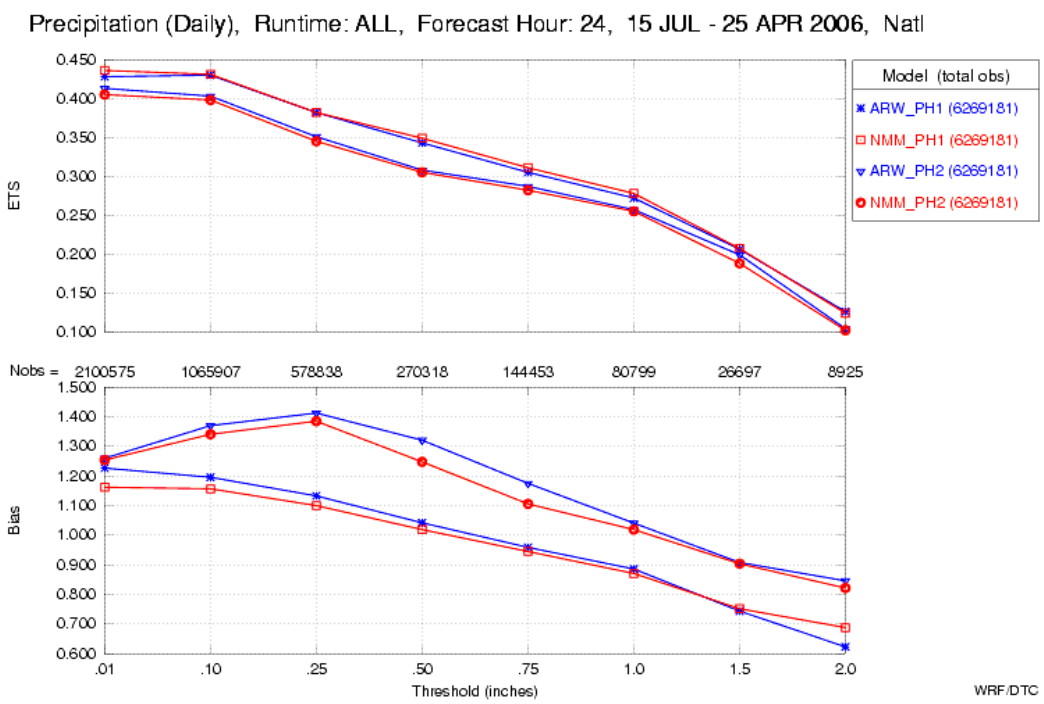


Figure 14. 24-h precipitation verification over all 4 seasons for 4 WRF model versions: ARW-Phase 1 physics, NMM-Phase 1, ARW-Phase 2, and NMM-Phase 2. Equitable threat score (ETS) on top and bias on bottom. (From WRF-DTC verification display at http://ruc.noaa.gov/wrf/RR/testing/NCEP_verif .)

B. **Visual evaluation**

In order to gauge qualitatively the differences between the forecasts produced by the ARW and the NMM, a subjective evaluation was made by individuals in GSD. The primary source material for this evaluation was the case study graphics web page set up by the DTC for this purpose: http://bolas.fsl.noaa.gov/mab/wrfrr .

For the initial time and every 3h out to 24h, this web site provided images of the following 2-d fields:

2-m temperature
2-m dew point
10-m wind speed with barbs
Total, convective and (for some days) grid-scale precipitation for the previous 3h
Run-accumulated precipitation
Snow-water equivalent on the ground
Precipitation type
Convective Available Potential Energy (CAPE) and Convective Inhibition (CIN)
Precipitable water vapor
Mean sea-level pressure
850-hPa height, temperature, relative humidity (RH) and wind (barbs and isotachs)
850 – 500-hPa mean RH
700-hPa height and wind barbs
700-hPa vertical motion (as omega = Dp/Dt)
500-hPa height and vorticity
250-hPa height, wind barbs and isotachs
For a few runs, plots of forecast soundings for KDVN, Davenport IA.

In addition to the individual images, it was also possible to display ARW and NMM images of the same fields side by side. Further, ARW – NMM difference fields were available for most of the scalar fields, including wind speed. A looping capability was added to loop from initial time to 24h by 3-h increments.

In the course of the evaluation, nearly all the available images were at least viewed as part of a loop, if not examined individually. Features that caught our attention were then explored more fully. We paid special attention to wind, temperature and humidity at the available pressure levels, 700-hPa vertical motion, and the components of the precipitation. We also used the statistical verification, particularly the time series plots of RMSE and bias for the individual runs, as a guide to particular forecasts or series of forecasts where differences between the cores were notable. In what follows we summarize our findings.

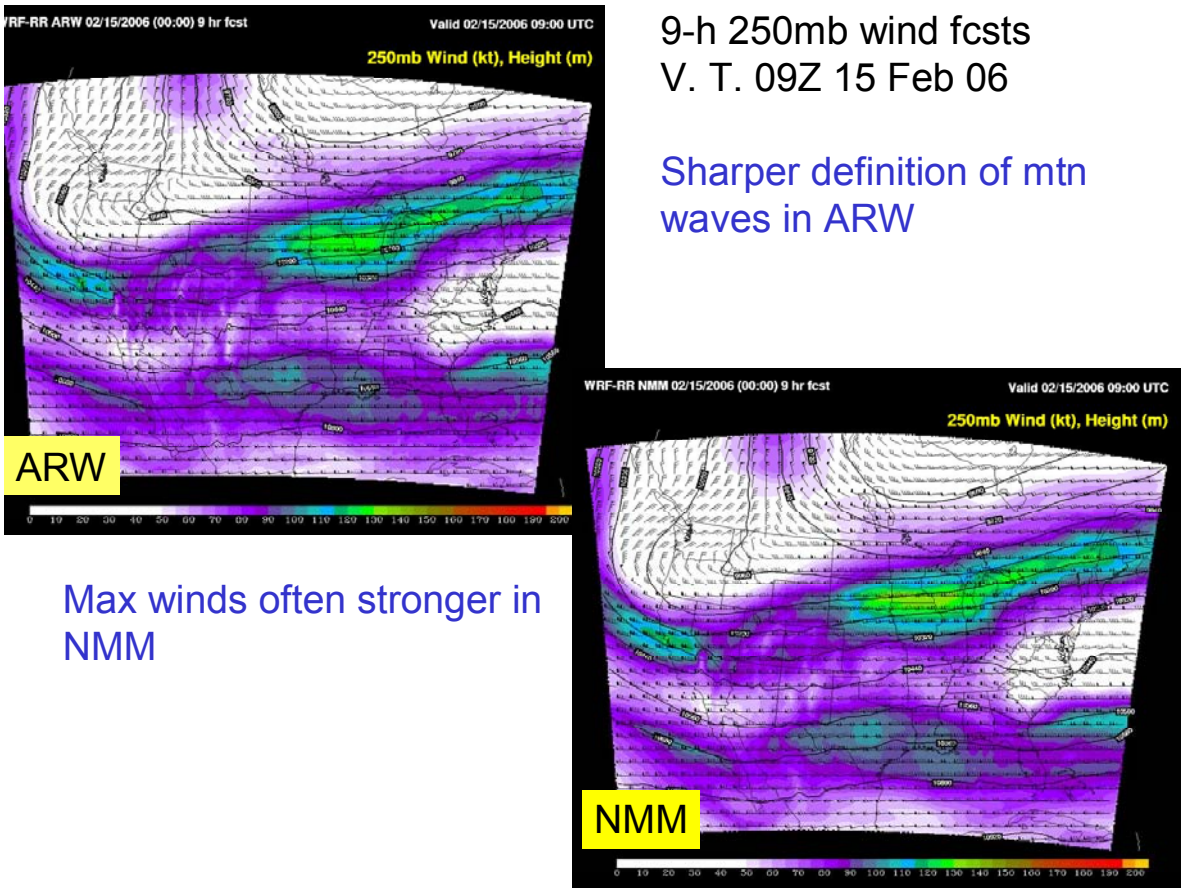
General impressions

Forecasts overall were very similar. Sometimes we had to look hard, and at small details, to see readily discernible differences.

As has been pointed out elsewhere, we might have expected this, given the constraints imposed in the core-test design. But, it is fair to say that, based on our experience in comparing ARW and NMM runs in other contexts and using much higher horizontal resolutions (Storm Prediction Center Spring Program evaluations in 2004 and 2005, see Kain et al 2006; the DTC Winter Forecast Experiment in 2005) we were a little surprised at how similar the forecasts from the core-test runs really were. To be fair to these other experiments, their stated purpose was not primarily to compare the performance of the ARW and NMM cores, as was this RR Core Test. Nevertheless, the point stands that to make focused, quantitative comparisons between different models, or between alternative formulations in the same model (e.g., numerical diffusion, physics packages), strict control on the experimental design is necessary.

The most obvious differences in all fields occurred at horizontal scales below 10-15 Δx, i.e., below ~150km. For the core configurations used for the full Phase 1 and Phase 2 runs (see DTC report) the ARW clearly exhibited more detail at these scales. This was most obvious in the horizontal wind fields and the omega fields. (See AWRP PDT reports from the In Flight Icing and Turbulence PDTs for other aspects of this difference.) Given that for our core configurations the flow should be essentially hydrostatic at all resolvable scales, this indicates that differences were tied primarily to prediction of the divergent component of the horizontal flow. It follows then that the differences are primarily affecting the gravity-wave modes at scales approaching the limit of accurate resolvability.

Examples of this are shown in Fig. 15. This shows 250-hPa wind barbs and isotachs, as well as geopotential heights, exhibiting sharper small-scale detail in the ARW as compared with NMM. Much of this difference is associated with mountain-wave activity, more prominent in the ARW than the NMM, produced by strong westerlies across the Rockies (e.g., CO and NM). The corresponding vertical motion (omega) field at 700 hPa can be seen in Fig. 16, where the NMM – ARW differences are more evident. These types of differences are typical, and are not confined to areas expected to have terrain-induced vertical motion.



9-h 250mb wind fcsts
V. T. 09Z 15 Feb 06

Sharper definition of mtn waves in ARW

Max winds often stronger in NMM

Figure 15. ARW and NMM 9-h 250-hPa wind (barbs, knots and isotachs, scale on bottom) and geopotential-height forecast for 0900 UTC 15 February 2006. Note especially speed variations over southwest US and jet streak upper west of Lake Michigan.

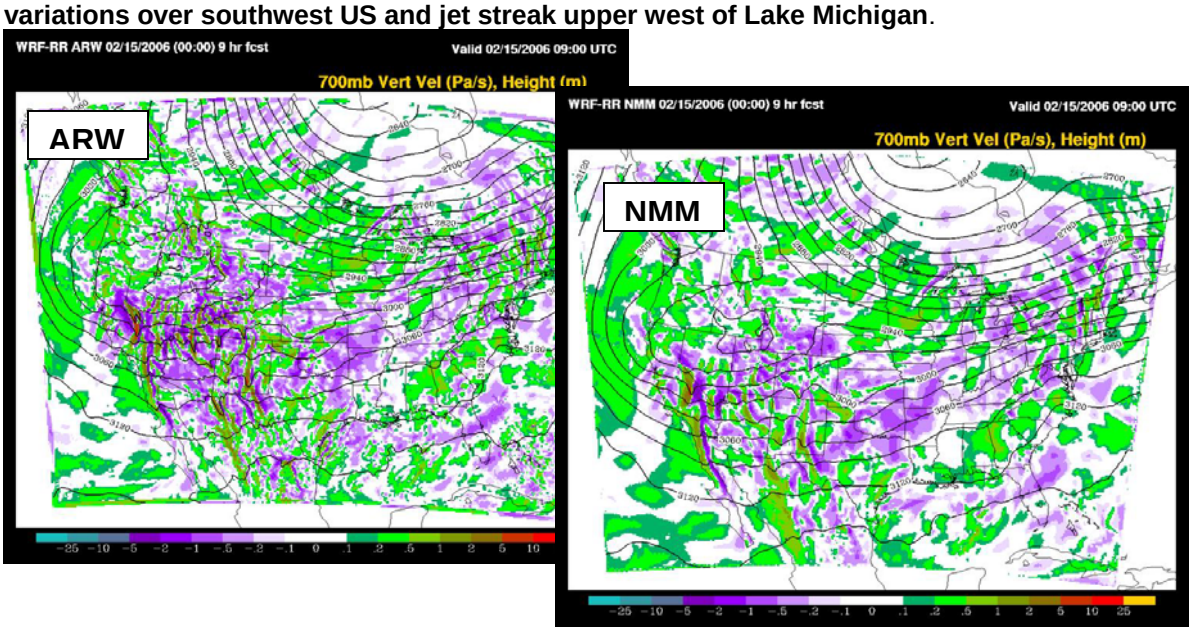


Figure 16. As with Fig 15 but for the vertical motion field (actually omega = Dp/Dt, Pa/s, so negative is upward motion). At this level, +1 pa/s roughly equals -10cm/s.

It should be noted that both cores produce a considerable amount of mountain-wave activity. It has not been possible to determine how much of the apparent larger amplitude mountain-wave response in the ARW can be attributed to the generally sharper ARW terrain (Fig. 2). We can say that the greater amount of small-scale detail in the ARW omega field away from terrain implies that sharper ARW terrain is not the sole contributor. Differences between the cores in the omega fields are discussed from different perspectives in the reports by the Turbulence and the In-Flight Icing PDTs.

This aspect of the comparison of the cores engendered considerable interest and dialogue during the evaluation phase of the core test. As a result, limited experiments were conducted with a small subset of runs that explore this issue a bit further (see App. E). Further, high-resolution (2-km horizontal grid spacing) runs over the southern Sierra in California are being conducted for the mountain-wave Turbulence and Rotors EXperiment (T-REX) using both cores and phase 1 physics (but without convective parameterization). These runs may shed further light on this matter as it pertains to mountain waves.

Aside from this, but not independent from it, the most obvious differences were in precipitation. There was a tendency for the ARW to produce slightly larger coverage of precipitation at the lighter amounts. This was perhaps most noticeable in post-frontal air masses over the Great Lakes and off the East Coast of the US. Consistent with this, the statistical verification of precipitation forecasts showed a higher bias at the small amounts for the ARW (Fig. 14). For the heavier amounts there was usually a correspondence between readily identifiable mesoscale heavier precipitation areas (say, in excess of 10mm/3h having areas of order 10^4 km^2) between the 2 cores. These areas usually at least partially overlapped. Differences between the cores were generally less than differences between the phases, particularly for precipitation from the convective parameterizations.

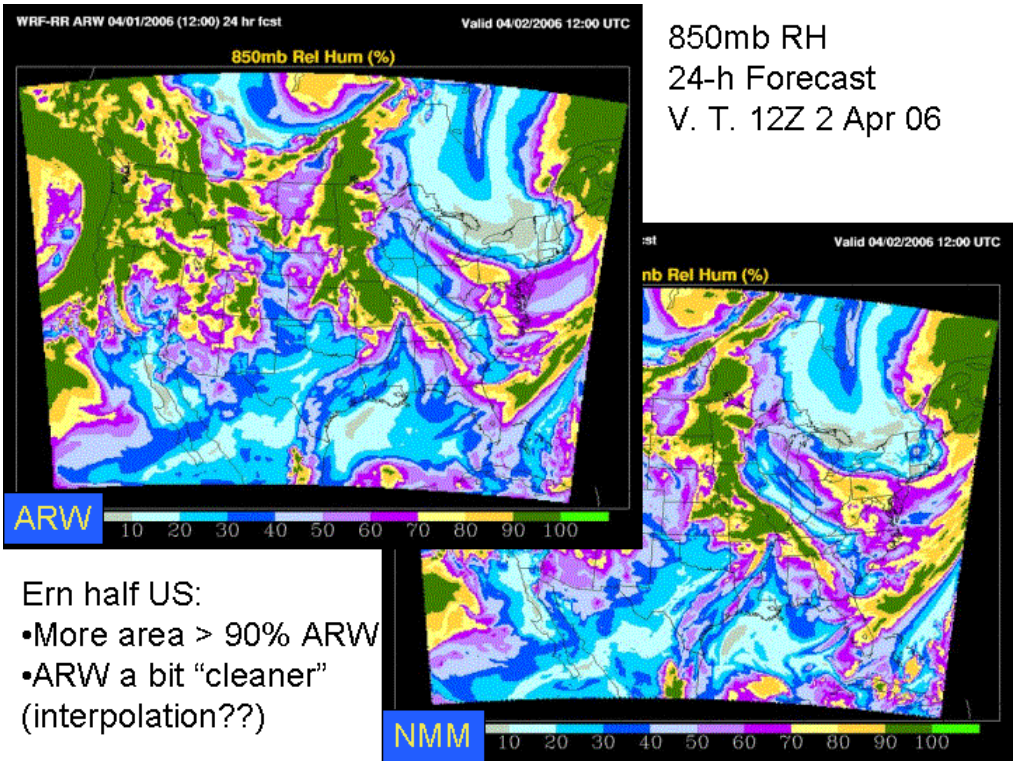


Figure 17. 24-h phase 2 forecast of 850-hPa relative humidity valid 1200 UTC 2 April 2006. Note the dark green areas over the upper Mississippi Valley referred to in the text.

Specific Impressions

The ARW had slightly larger orographic precipitation enhancement in the western US than the NMM. As noted in Appendix E, this may be largely attributable to the horizontal divergence damping in the NMM. Terrain differences also clearly contributed.

East of the Rockies, the ARW tended to produce larger, more continuous areas of relative humidity > 90% at 850-hPa in cold-season warm-advection situations than did the NMM, for both Phase 1 and 2 physics. An example of this is shown in Fig. 17, where the dark green areas are > 90%. This behavior was cited by the Inflight Icing PDT (Ben Bernstein, June 2006, personal communication) as an advantage of the ARW over the NMM as input to their Forecast Icing Product, and was part of their basis for recommending the ARW core. We have no obvious explanation for this behavior, which was not seen in every such case.

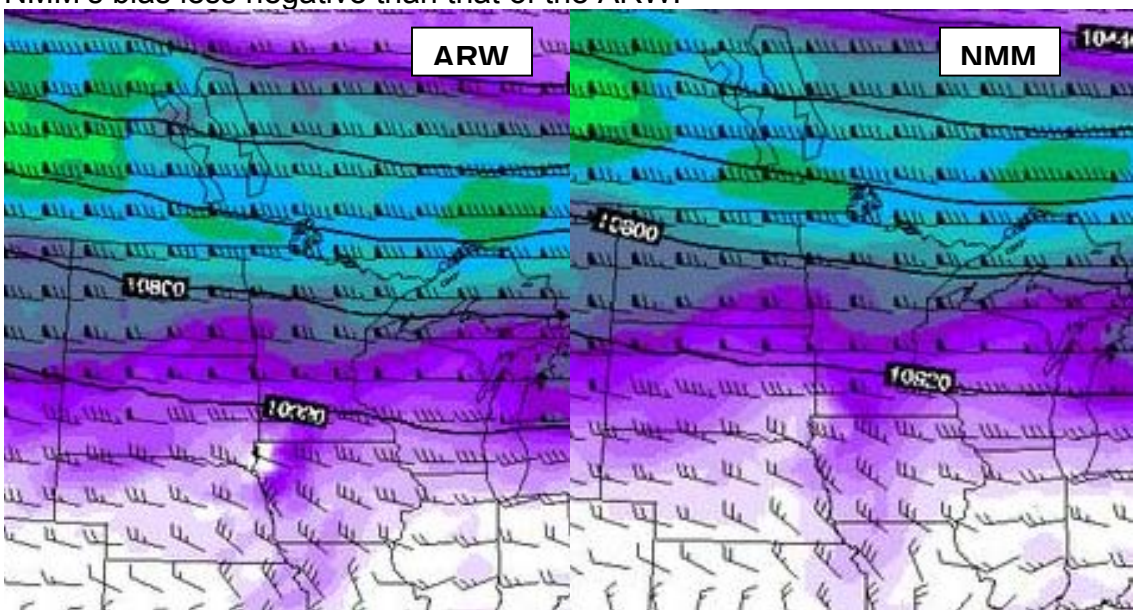
The ARW often produces slightly stronger 10m AGL winds than the NMM. This corresponds to a similar bias noted at 850 hPa in the statistical verification (Fig. 10). As noted in section 3.C under "terrain elevation specification", we went to considerable effort to ensure the roughness length for momentum was the same in both cores, so we believe this difference is intrinsic to the core configurations we used.

We also noticed that cold fronts in the ARW often moved very slightly faster (no more than 2-3%) than in the NMM. This behavior seems consistent with the slightly stronger low-level winds in the ARW noted above.

The ARW exhibits a more rapid resolvable-scale response to convection than the NMM. This was particularly noticeable with the BMJ convection in phase 1. This scheme is known to produce notable perturbations in the mass and wind fields as a result of its adjustment process (Bukovsky et al 2006, see especially their Fig. 8). What we noticed is that the response seemed to be initially larger and to become evident more quickly in the ARW than the NMM for a given intensity of convective precipitation. An example of this is shown in Fig. 18 for a well-forecast nocturnal convective case near the South Dakota – Iowa border in the two forecasts.

Note the small, strong perturbation in the 250-hPa winds in NW IA for the ARW. A broader perturbation is present in the same area in the NMM forecast. Both forecasts had about the same intensity and areal coverage of convective precipitation and little grid-scale precipitation. This difference in response between ARW and NMM could arguably be a consequence of the horizontal divergence damping in the NMM.

On the other hand, peak winds in the core of strong jet streaks at 250-hPa were usually forecast to be stronger in the NMM than the ARW. For example, a census of phase 2 winter 12-h forecasts having jet streaks with peak wind of at least 150 kt in at least one of the cores showed the NMM to clearly have a stronger peak wind speed 9 times, the ARW only twice, and too close to call with confidence, 19 times. This is consistent with the statistical verification that shows both models to have a low wind-speed bias at this level (Fig. 4b), but with the NMM's bias less negative than that of the ARW.



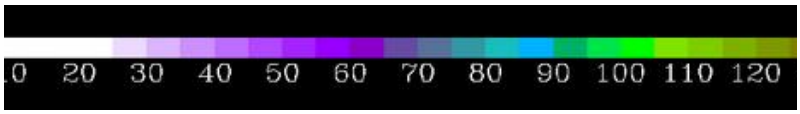


Figure 18. 9-h phase 1 forecast of 250-hPa isotachs (kts) and wind barbs at 0900 UTC 21 July 2005, for both ARW and NMM. The scale for the isotachs is shown at the bottom.

A curious phase-1 result was that the NMM typically had more convective precipitation than the ARW. Again, this was not seen in every situation, and was not apparent in the phase-2 comparisons. Systematic phase-2 precipitation differences resulting from the Grell-Devenyi scheme were most evident in regions of weak convective forcing but appreciable CAPE, for example the Southeast in summer: the ARW systematically produced more localized (probably one or 2 grid points in size) areas of convective precipitation > 3.5 mm/h. An example of this can be seen in Figure 5 of the report by the Convective Weather PDT. It was apparent also in the 3-h convection precipitation plots on the web site.

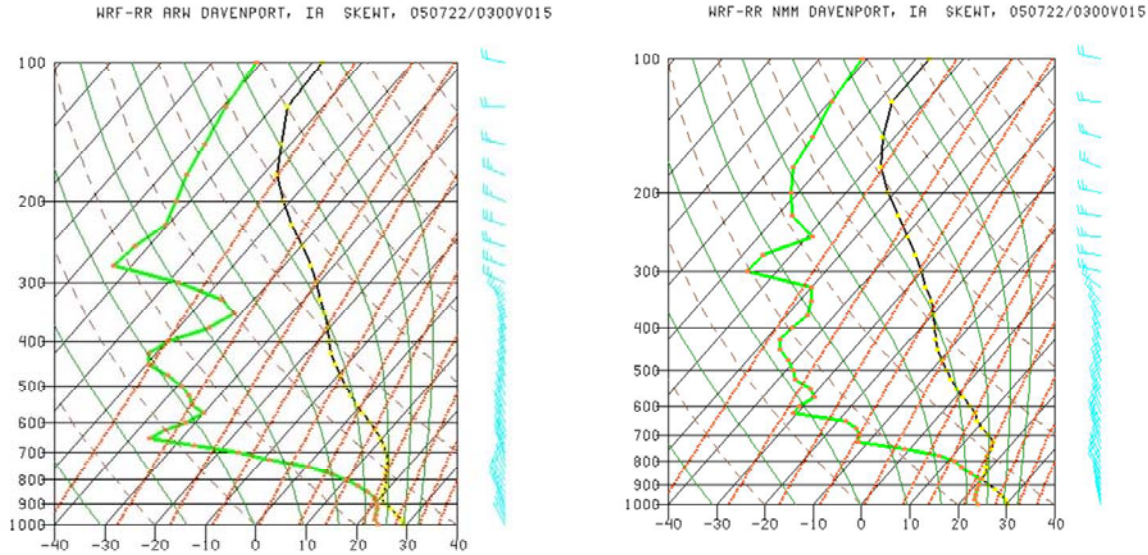


Figure 19. Skew-T plots from the location of KDVN, Davenport IA, for a 15-h phase 2 forecast initialized at 1200 UTC 21 July 2005. The ARW forecast sounding is on the left, NMM on the right.

Finally, it is apparent that forecast vertical soundings by the NMM show slightly more realistic structure than the ARW. Particularly, the mixed layer and the transition layer above are more often more realistically portrayed in the NMM forecasts, as shown in Fig. 19. Differences here are subtle, but the NMM rendition of this mixed layer shows a slightly more uniform theta and water-vapor mixing ratio profile than the ARW. The transition at the top of the mixed layer is a bit sharper in NMM also. This type of difference was first noted in T-REX soundings. This type of difference is not apparent in all mixed layer situations. Because the Storm Prediction Center uses model-predicted soundings, particularly from the RUC, as an important guidance element, this difference,

though slight, has motivated a supplemental experiment that is discussed elsewhere (see Appendix E).

Discussion

The core-test ARW and NMM forecasts were very similar, but systematic behavior differences were apparent, and broadly consistent with the statistical verification when comparison was warranted. Some of these differences favor the ARW and some the NMM.

The increased detail at small scales in the ARW relative to the NMM is ambiguous in its bearing on the decision between the cores, in our opinion. The only firm piece of evidence we have that this detail is meaningful, at least in a statistical sense, is the improved performance of the ARW over the NMM in the statistically based turbulence diagnostic as discussed by the Turbulence PDT in their report. On the other hand, forecasters using an ARW version of the RR as guidance may find the additional detail in its forecasts troublesome unless they can have more confidence that it is meaningful.

The systematic behavioral differences we observed that in our judgment favor the NMM are as follows.

- More realistic sounding structure (though differences are subtle and not always favoring the NMM)
- Stronger peak speed in upper level jet streaks (consistent with statistical verification showing NMM has less negative bias in wind speeds at 250-hPa)
- Less small-scale detail that may make NMM advantageous for subjective forecaster interpretation
- Slightly better wind forecasts near the surface

Systematic behavioral differences that in our judgment favor the ARW are as follows.

- Less explicit damping of any meteorologically significant gravity waves and other small-scale phenomena allowing them to more realistically contribute to model prediction of significant weather, including turbulence.
- More realistic prediction of relative humidity (and, by implication clouds) in warm-advection regions.
- More realistic precipitation patterns, particularly in regions of convection.
- Slightly better temperature forecasts at low levels

C. NMM/ARW experimental results from July 2006

See Appendix E.

6. Summary on core-test preferences from WRF-RR core-test reports other than GSD

See Appendix D for more detailed summaries for all reports below, including results beyond core preferences.

- Turbulence PDT

Both structure function analysis and validation of turbulence prediction against PIREPs favor ARW over NMM.

- Convective weather PDT

Occurrence of higher rainfall rates and subjective evaluation of convection forecasts favor ARW.

- In-flight icing PDT

Evaluation of forecast icing potential (FIP) with both cores shows no clear difference. Higher effective resolution of ARW shows it has potential for more precise forecasts than NMM.

- Ceiling/visibility PDT

Both cores performed similarly; no preference expressed.

- DTC

Reviewed statistical results and summarized showing a few areas of ARW advantage, differences in bias, and many areas of no difference. DTC did not express preference.

- Aviation Weather Center

Preferred ARW by reason of increased model diversity since NAM uses NMM.

- Storm Prediction Center

No preference expressed.

7. Discussion and conclusions

A. Our recommendation

Aviation and severe weather forecasting interests have specific requirements that the Rapid Refresh must address related to forecasts of icing, turbulence, ceiling/visibility, convection, winds (at all levels) and temperature/moisture stability profiles in the convective weather environment. With this focus:

- Based on the recent 9-month evaluation, we in the Rapid Refresh development group in NOAA/ESRL/GSD recommend, by a slight margin, the ARW core over the NMM core for the initial operational Rapid Refresh implementation planned for 2008.
- Some significant advantages were evident for one core or the other, dependent on variable or vertical level, with a slight edge for ARW overall, but we judged that there was no strong overall advantage for either.
- Therefore, we will fully support your decision, regardless of which core you choose, but please consider carefully the results presented earlier in this report.

- This recommendation is based on an extensive evaluation of core-test forecasts by the organizations below and their combined recommendations:
    - NOAA/ESRL/GSD
    - Developmental Testbed Center (NCAR and NOAA/ESRL/GSD)
    - Product Development Teams for aviation weather products funded through the FAA Aviation Weather Research Program

- Our recommendation is for the *initial* operational implementation of Rapid Refresh currently planned for 2008. NOAA's purpose for use of WRF in NCEP operational models is to accelerate transition from research into operations on an ongoing basis. Yearly update packages will be made to the Rapid Refresh code, similar to the process for the current Rapid Update Cycle (RUC) model and other NCEP operational forecast models. Given the likelihood that important further developments will emerge from the WRF modeling community, we recommend that a re-evaluation of the WRF-RR dynamic core should be conducted every 2 years based on WRF community developments.

- As a result of the WRF-RR core-test evaluation, a full set of physical parameterizations from two different suites are now available for use with either the NMM or ARW cores, including the parameterizations likely required for the operational Rapid Refresh. This outcome is obviously beneficial to the Rapid Refresh development project as well as for the WRF community at large.

B. Key points for decision from aviation and severe weather NWP point of view

1) ARW advantages

Major advantages

- Upper-level wind. This is apparent in aircraft verification. Rawinsonde verification (where ARW advantage was even stronger) is considered flawed. (see section 5.A.1)
- Lower-troposphere temperature (see section 5.A.2)
- Lower-troposphere relative humidity, primarily at 850 hPa, considered to be potentially important for icing and ceiling forecasts. (See section 5.A.3.)
- Turbulence (see objective verification results in Appendix D).

Secondary advantages

- Community involvement – Currently much more significant with ARW testing and applications than with NMM. This may be different in 1 year after additional NMM community exposure.

Example: NCAR is working on improving the ARW digital filter initialization (DFI), which is required for the Rapid Refresh to allow sufficiently quiet 1-h forecasts.

2) NMM advantages

Major advantages

- Wind speed bias, particularly at upper levels.
- Precipitation bias

Secondary advantages

- Code already developed for positive definite horizontal advection and calling microphysics less often than every dynamic time step.
- NCEP/EMC will continue to develop NMM in the context of the NAM application.

3) Model characteristics that we did not consider as factors in the decision

- Speed of dynamic core.
    - Higher speed will allow higher resolution, assuming that the "effective resolution" is also equal.

The NMM allows (and even requires) the rotated Mercator grid, which has more nearly equal map-scale factor than other projections, including the Lambert conformal map projection. The ARW does not allow the rotated Mercator map projection. This factor may be significant with the Rapid Refresh domain, extending from near the equator to beyond 70°N.

Based on GSD timing tests between ARW and NMM cores, we currently estimate the NMM as 30-40% faster including equalizing horizontal advection computation between NMM and ARW, map-scale factor, and IJK-computation nearing availability with the NMM. This would be equivalent to ~10% in horizontal grid spacing.

- Improved NMM performance with 15 August 2006 changes to operational NAM-WRF-NMM.

The doubling of the horizontal diffusion coefficient with the operational NAM-NMM at NCEP on 15 August 2006 probably lowered forecast errors, according to NCEP presentations to GSD on 17 July. A similar modification could be made to the ARW with similar results. We note that improved verification scores against upper-air observations can also be obtained by smoothing output fields (as done in the RUC), and this is an alternative for either dynamic core without increasing internal diffusion or damping.

- Difference in smoothness of forecast grids

We considered the degree of smoothness as not an inherent advantage for either core and not a factor in our recommendation.

- 700-hPa temperatures with Phase 1 physics.

Phase 1 experiments showed a significant advantage for the NMM model in 700-hPa temperatures. This advantage disappeared using Phase 2 physics. Moreover, 700-hPa temperature forecasts were somewhat more accurate with Phase 2 physics, so we disregarded the Phase 1 temperature results at 700 hPa.

Acknowledgments

The WRF-Rapid Refresh core test evaluation has been dependent on a wide group of participants from GSD, DTC, and NCAR beyond those listed in Appendix A for WRF model code development. We note, in particular, Tanya Smirnova (GSD) and Ligia Bernardet (DTC, GSD). Betsy Weatherhead and Greg Noonan (ESRL/GMD and DTC) provided hundreds of graphs for consideration, many extracted for section 5. Moreover, they performed the statistical significance testing. The RTVS/DTC/WRF interactive statistics web site by Andy Loughe (GSD, DTC) allowed most of the ongoing evaluation as experiments were performed. Randy Collander (GSD, DTC) developed the forecast graphics and an associated web site. Barry Schwartz (Schwartz Weather Consultants) performed verification and significance testing. Louisa Nance (DTC, NCAR) was a vital colleague for this study. Steve Koch (DTC, GSD) and Bob Gall (DTC, NCAR) provided very helpful guidance in coordinating DTC activities. Tom Schlatter and Steve Weygandt provided excellent reviews of this report and helped to decide on the recommendations.

Glossary of acronyms

ARW	Advanced Research WRF (WRF dynamic core)
AWRP	Aviation Weather Research Program (FAA-funded)
CONUS	continental US (lower 48 states)
DTC	Developmental Testbed Center, run jointly by NCAR and NOAA
EMC	NCEP Environmental Modeling Center

ESRL	Earth System Research Laboratory in NOAA Research
GMD	Global Monitoring Division in NOAA/ESRL
GSD	Global Systems Division in NOAA/ESRL
LSM	Land-surface model (multi-layer soil/vegetation/snow parameterization)
NAM	North American Mesoscale operational model running at NCEP (running with WRF model and NMM dynamic core as of 20Jun06)
NCEP	NOAA National Centers for Environmental Prediction
NMM	Non-hydrostatic Mesoscale Model (WRF dynamic core)
PBL	Planetary boundary layer
PDT	Product development team (under FAA-sponsored Aviation Weather Research) Program)
RAOB	Rawinsonde
RR	Rapid Refresh
RTVS	Real-Time Verification System (developed at GSD Aviation Branch)
RUC	Rapid Update Cycle
SST	Sea-surface temperature

References

Benjamin, S.G., D. Devenyi, S.S. Weygandt, K.J. Brundage, J.M. Brown, G.A. Grell, D. Kim, B.E. Schwartz, T.G. Smirnova, T.L. Smith, and G.S. Manikin, 2004: An hourly assimilation/forecast cycle: The RUC. *Mon. Wea. Rev.*, **132**, 495-518 (Feb. issue).

Bukovsky, M.S., J. S. Kain and M. E. Baldwin, 2006: Bowing convective systems in a popular operational model: Are they for real? *Wea. Forecasting*, **21**, 307-324.

Kain, J.S., S. J. Weiss, J. J. Levit, M. E. Baldwin, D. A. Bright, 2006: Examination of convective-allowing configurations of the WRF model for prediction of severe convective weather: the SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167-181.

Lindborg, E., 1999: Can the atmospheric kinetic energy spectrum be explained by two-dimensional turbulence? *J. Fluid Mech.*, **388**, 259-288.

Sharman, R., C. Tibaldi, G. Wiener, and J. Wolff, 2006: An integrated approach to mid- and upper-level turbulence forecasting. *Wea. Forecasting*, **21**, 268-287.

WRF-Rapid Refresh Core-Test Reports

Benjamin, S., and J.M. Brown 2006: WRF Rapid Refresh dynamic core evaluation and recommendations, [this report], 62 pp.

Bernstein, B., 2006: In-flight Icing PDT WRF results, 14 pp.

Nance, L.B., 2006, Weather Research and Forecasting Core Test DTC Report, 36 pp.

Pinto, J., and C. Phillips, 2006: Convective Weather PDT analysis of the impact of WRF-RR cores on current and planned convective weather products, 12 pp.

Sharman, R., and R. Freilich, 2006: Turbulence PDT WRF ARW/NMM core comparisons, 13 pp.

Simard, J., P. Herzegh, and R. Bateman, 2006: WRF core test ceiling and visibility performance for the ARW and NMM dynamic cores, 6 pp.

Weatherhead, E., G. Noonan, T. Fowler, L. Bernardet, L. Nance, and S. Koch, 2006: Summaries of statistical analyses of differences in relative humidity, temperature, wind and precipitation between ARW and NMM WRF cores using both sonde and aircraft data, 7 pp.

Key web sites

Interactive WRF-RR interactive web site –
http://ruc.noaa.gov/wrf/RR/testing/NCEP_verif

Graphics for ARW and NMM forecasts from WRF-RR core-test comparison
http://bolas.fsl.noaa.gov/mab/wrfrr/

Appendix A: Modifications to WRF model necessary for the Core Test Pertaining to Physics Schemes

A primary goal of the core-test preparation was to ensure, via whatever code changes necessary, that the physical parameterizations were available and ran identically in both cores for each phase. Tanya Smirnova of GSD played a pivotal role in this difficult and often tedious process. Others at NCAR/MMM and NCAR/RAL and the DTC as well as GSD and NCEP/EMC were also closely involved in many aspects. In this section, we supplement the material shown in Table 2 of this document by summarizing this preparation as it affected the physics schemes in both phases.

<u>Phase 1</u>

A major decision near the beginning of the core-test preparations (October-November 2005) concerned which version of the Noah land-surface scheme to adapt. The most current version of Noah, option 2 in the namelist.input file, at that time only ran in the ARW. The NCEP was using "option 99" in its version of the NMM being prepared for the North American Mesoscale implementation scheduled for 2006. This version, however, was not checked out with ARW v2.1. GSD looked into this issue carefully and concluded that, although there were serious plans to combine option 2 with option 99 to make a new option 2 that would be implemented in the NMM as the unified Noah LSM, this would not occur until after the NAM operational implementation, too late for phase 1 of the core test. The easier route appeared to be to implement the NAM-NMM version of "option 99" fully into ARW. This was accomplished in December 2005.

The WRF v 2.1.1 release in November 2005 had the Ferrier microphysics implemented in the ARW with hydrometeor coupling to the GFDL radiation scheme from Brad Ferrier (NCEP/EMC) and Jimy Dudhia (NCAR/MMM). A few further modifications were necessary to ensure the proper attenuation of short-wave radiation in the GFDL radiation scheme for ARW when convection was active.

A more time-consuming issue was proper initialization of the RUC hydrometeors for use in the Ferrier scheme in both the ARW and NMM Standard Initializations. Since the RUC hydrometeors came from the (2004 version) of the NCAR microphysics, it was necessary to introduce code into real.exe to initialize the total condensate advected by the Ferrier scheme, partition it into liquid and ice, and estimate a riming fraction based on assumed riming fractions of the three ice species and the proportion of mixing ratios of graupel, snow and ice in the RUC initial conditions.

On 30 January 2006 WRF v2.1.2 was released to the community. Motivated by a desire to keep the core-test code as current and as repository-compliant as

practical, earlier core-test mods to v2.1.1 were merged with this new release. This revealed some subtle compile difficulties on ijet that were finally tracked down and fixed.

On 17 February 2006, NCEP/EMC announced extensive changes to a number of physics-related modules (as well as some other changes) to the NMM. Since this was the code that would be included in the initial WRF-NAM implementation scheduled for June 2006, NCEP/EMC requested that this NMM code be used for the core test. A number of platform specific compiler issues also arose with the testing of this code, as well as discovery of 2 egregious bugs that were adversely affecting NMM forecast performance. Fixes relevant to the WRF-NAM were then passed back to NCEP. Further, with these NCEP modifications as they were being applied in the core test, the roughness length for momentum was found to be larger by a factor of 2 or more in some areas in the NMM over the ARW. To eliminate this inconsistency, a compromise was made to remove all terrain dependency in the roughness length (NMM) and add a uniform value of 10cm to all land points (ARW).

Phase 2

Referring to Table 2 in this document, it was necessary to implement both the RUC LSM and the Grell-Devenyi (GD) convection scheme in the NMM. Further, the version of the GD scheme available through the namelist in the ARW dated back about 4 years. Georg Grell (GSD) updated this in the ARW to the latest version of his scheme, which is now used in the operational RUC. Tanya Smirnova implemented the RUC LSM in the NMM and assisted Georg in implementing the GD scheme. No serious unforeseen problems arose in this process.

Greg Thompson of NCAR/RAL developed a new version of the NCAR microphysics (Thompson et al 2004) that was already in the WRF v2.1 release of July 2005. Since this new version seemed more likely to be used in the operational RR than the 2004 version, it was decided by the core-test team to implement it in both ARW and NMM. (It was originally implemented and checked out in the MM5 model.) Greg Thompson did the ARW implementation, with assistance from others at NCAR/MMM and GSD, and this was checked out before he and Tanya accomplished the same for the NMM. In each case, there were special challenges in the feedbacks between the microphysics and both the GFDL radiation and the Mellor-Yamada-Janjic PBL (MYJPBL), both resulting from features in the February 2006 NCEP versions of these schemes that hardwired them to the Ferrier microphysics. In order to make the MYJPBL behave identically in both cores, the February 2006 NCEP version of the MYJPBL (see above) was altered to mix only water vapor, but no hydrometeors.

Finally, a patch to zero out very small hydrometeor values in WRFpost was required to prevent the grib encoder from crashing.

Appendix B.1: WRF-ARW namelist used for Phase 2

```
 &time_control
 RUN_DAYS = 1,
 RUN_HOURS = 0,
 RUN_MINUTES = 0,
 RUN_SECONDS = 0,
 START_YEAR = 2005,
 START_MONTH = 11,
 START_DAY = 04,
 START_HOUR = 12,
 START_MINUTE = 00,
 START_SECOND = 00,
 END_YEAR = 2005,
 END_MONTH = 11,
 END_DAY = 05,
 END_HOUR = 12,
 END_MINUTE = 00,
 END_SECOND = 00,
 INTERVAL_SECONDS = 10800,
 INPUT_FROM_FILE = .true.,.true.,.true.,
 HISTORY_INTERVAL = 180, 60, 60,
 FRAMES_PER_OUTFILE = 1, 1, 1,
 RESTART = .false.,
 RESTART_INTERVAL = 10000,
 IO_FORM_HISTORY = 2
 IO_FORM_RESTART = 2
 IO_FORM_INPUT = 2
 IO_FORM_BOUNDARY = 2
 DEBUG_LEVEL = 0
/
 &domains
 TIME_STEP = 60,
 TIME_STEP_FRACT_NUM = 0,
 TIME_STEP_FRACT_DEN = 1,
 max_dom = 1
 s_we = 1,
 e_we = 400,
 s_sn = 1,
 e_sn = 304,
 s_vert = 1,
 e_vert = 50,
 dx = 13300.0,
 dy = 13300.0,
 grid_id        = 1,
 parent_id      = 1,
 i_parent_start = 0,
 j_parent_start = 0,
 parent_grid_ratio = 1,
 parent_time_step_ratio = 1,
 FEEDBACK = 1,
 SMOOTH_OPTION = 0
/
 &physics
 MP_PHYSICS = 8,     5,     5,
 mp_zero_out = 2,
```

```
 mp_zero_out_thresh = 1.e-12
 RA_LW_PHYSICS = 99,     1,     1,
 RA_SW_PHYSICS = 99,     1,     1,
 CO2TF = 1,
 RADT = 30,    10,    10,
 SF_SFCLAY_PHYSICS = 2,     1,     1,
 SF_SURFACE_PHYSICS = 3,     1,     1,
 BL_PBL_PHYSICS = 2,     1,     1,
 BLDT = 0,     0,     0,
 CU_PHYSICS = 3,     0,     0,
 CUDT = 0,     5,     5,
 ISFFLX = 1,
 IFSNOW = 1,
 ICLOUD = 1,
 SURFACE_INPUT_SOURCE = 1,
 NUM_SOIL_LAYERS = 6,
 MAXIENS = 1,
 MAXENS = 3,
 MAXENS2 = 3,
 MAXENS3 = 16,
 ENSDIM = 144,
/
 &dynamics
 DYN_OPT = 2,
 RK_ORD = 3,
 W_DAMPING = 1,
 DIFF_OPT = 1,
 KM_OPT = 4,
 DAMP_OPT = 1,
 ZDAMP = 5000.,  5000.,  5000.,
 DAMPCOEF = 0.02,
 KHDIF = 0,      0,      0,
 KVDIF = 0,      0,      0,
 SMDIV = 0.1,    0.1,    0.1,
 EMDIV = 0.01,   0.01,   0.01,
 EPSSM = 0.1,    0.1,    0.1
 NON_HYDROSTATIC = .true., .true., .true.,
 TIME_STEP_SOUND = 4,      4,      4,
 H_MOM_ADV_ORDER = 5,      5,      5,
 V_MOM_ADV_ORDER = 3,      3,      3,
 H_SCA_ADV_ORDER = 5,      5,      5,
 V_SCA_ADV_ORDER = 3,      3,      3,
/
 &bdy_control
 SPEC_BDY_WIDTH = 5,
 SPEC_ZONE = 1,
 RELAX_ZONE = 4,
 SPECIFIED = .true., .true.,.true.,
 PERIODIC_X = .false.,.false.,.false.,
 SYMMETRIC_XS = .false.,.false.,.false.,
 SYMMETRIC_XE = .false.,.false.,.false.,
 OPEN_XS = .false.,.false.,.false.,
 OPEN_XE = .false.,.false.,.false.,
 PERIODIC_Y = .false.,.false.,.false.,
 SYMMETRIC_YS = .false.,.false.,.false.,
 SYMMETRIC_YE = .false.,.false.,.false.,
 OPEN_YS = .false.,.false.,.false.,
```

```
 OPEN_YE = .false.,.false.,.false.,
 NESTED = .false., .true., .true.,
/
 &namelist_quilt
 NIO_TASKS_PER_GROUP = 0,
 NIO_GROUPS = 1,
/
```

Appendix B.2: WRF-NMM namelist used for Phase 2

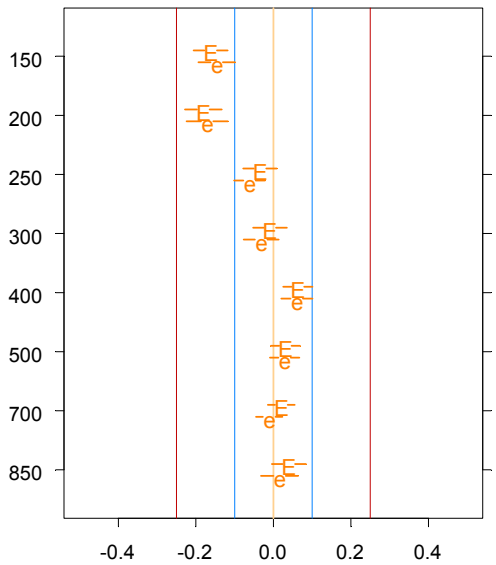
```
 &time_control
 RUN_DAYS = 1,
 RUN_HOURS = 0,
 RUN_MINUTES = 0,
 RUN_SECONDS = 0,
 START_YEAR = 2006,
 START_MONTH = 03,
 START_DAY = 29,
 START_HOUR = 00,
 START_MINUTE = 00,
 START_SECOND = 00,
 TSTART = 00,
 END_YEAR = 2006,
 END_MONTH = 03,
 END_DAY = 30,
 END_HOUR = 00,
 END_MINUTE = 00,
 END_SECOND = 00,
 INTERVAL_SECONDS = 10800,
 HISTORY_INTERVAL = 180,
 FRAMES_PER_OUTFILE = 1,
 RESTART = .false.,
 RESTART_INTERVAL = 6000,
 IO_FORM_HISTORY = 2
 IO_FORM_RESTART = 2
 IO_FORM_INPUT = 2
 IO_FORM_BOUNDARY = 2
 DEBUG_LEVEL = 0
/
 &domains
 TIME_STEP = 30,
 TIME_STEP_FRACT_NUM = 0,
 TIME_STEP_FRACT_DEN = 1,
 MAX_DOM = 1,
 S_WE = 1,
 E_WE = 281,
 S_SN = 1,
 E_SN = 436,
 S_VERT = 1,
 E_VERT = 50,
 DX = 0.08670,
 DY = 0.08380,
 GRID_ID = 1,
 TILE_SZ_X = 0,
 TILE_SZ_Y = 0,
 NUMTILES = 1,
/
 &physics
 MP_PHYSICS = 8,
 mp_zero_out = 2,
 mp_zero_out_thresh = 1.e-12
 RA_LW_PHYSICS = 99,
 RA_SW_PHYSICS = 99,
 RADT = 30,
```

```
 NRADS = 60,
 NRADL = 60,
 CO2TF = 1,
 SF_SFCLAY_PHYSICS = 2,
 SF_SURFACE_PHYSICS = 3,
 BL_PBL_PHYSICS = 2,
 BLDT = 0,
 NPHS = 2,
 CU_PHYSICS = 3,
 CUDT = 0,
 NCNVC = 2,
 tprec = 6,
 theat = 6,
 tclod = 6,
 trdsw = 6,
 trdlw = 6,
 tsrfc = 6,
 pcpflg = .false.,
 ISFFLX = 1,
 IFSNOW = 1,
 ICLOUD = 1,
 NUM_SOIL_LAYERS = 6,
/
 &dynamics
 DYN_OPT = 4,
 RK_ORD = 3,
 DIFF_OPT = 0,
 KM_OPT = 1,
 DAMP_OPT = 1,
 ZDAMP = 5000.,
 DAMPCOEF = 0.01,
 KHDIF = 0,
 KVDIF = 0,
 MIX_CR_LEN = 200.,
 SMDIV = 0.1,
 EMDIV = 0.01,
 EPSSM = 0.1,
 TIME_STEP_SOUND = 4,
/
 &bdy_control
 SPEC_BDY_WIDTH = 1,
 SPEC_ZONE = 1,
 RELAX_ZONE = 4,
 SPECIFIED = .true.,
 PERIODIC_X = .false.,
 SYMMETRIC_XS = .false.,
 SYMMETRIC_XE = .false.,
 OPEN_XS = .false.,
 OPEN_XE = .false.,
 PERIODIC_Y = .false.,
 SYMMETRIC_YS = .false.,
 SYMMETRIC_YE = .false.,
 OPEN_YS = .false.,
 OPEN_YE = .false.,
 NESTED = .false.,
/
 &namelist_quilt
```

```
 NIO_TASKS_PER_GROUP = 0,
 NIO_GROUPS = 1,
/
```

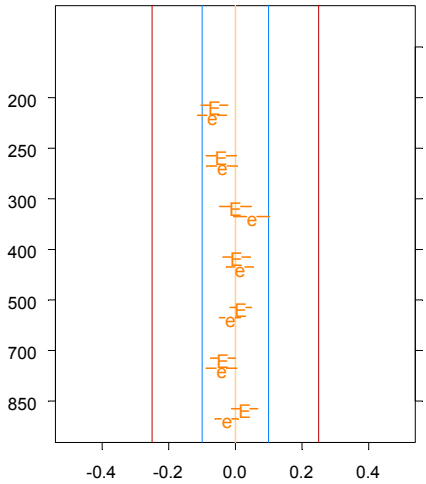
Appendix C – Annual summary statistics – 24-h forecasts, eastern verification area (to supplement figures shown in section 4.A)

Wind rmse Diff. 24h (ARW-NMM) Sondes: CONUS East annual



Capital letters are for phase 1; Weatherhead and Noonan

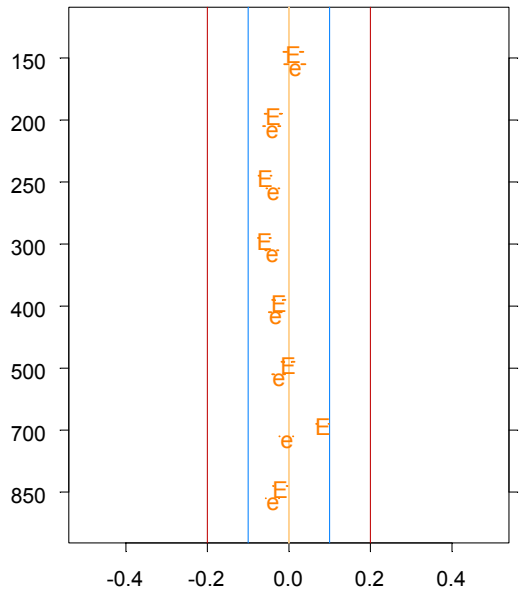
Wind rmse Diff. 24h (ARW-NMM) Aircraft: CONUS East annual



Capital letters are for phase 1; Weatherhead and Noonan

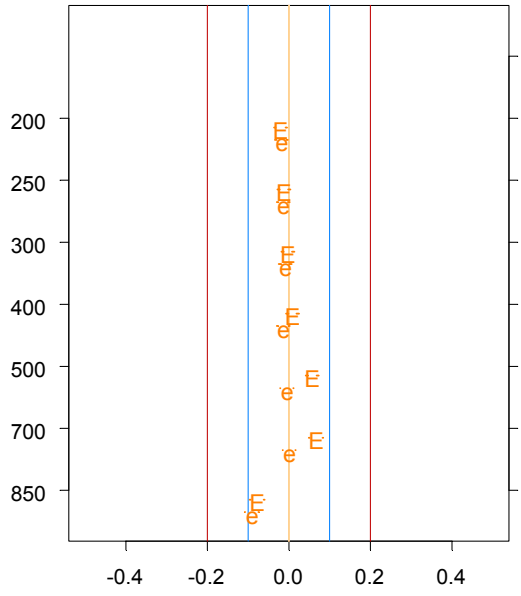
Figure C.1. Wind RMS vector error difference (ARW-NMM) for 24-h over eastern verification. Top – rawinsonde verification, botton, aircraft verification.

Temp. rmse Diff. 24h (ARW-NMM) Sondes: CONUS East annual



Capital letters are for phase 1; Weatherhead and Noonan

Temp. rmse Diff. 24h (ARW-NMM) Aircraft: CONUS East annual



Capital letters are for phase 1; Weatherhead and Noonan

Figure C.2. Temperature RMS error difference (ARW-NMM) for 24-h over eastern verification. Top – rawinsonde verification, botton, aircraft verification.

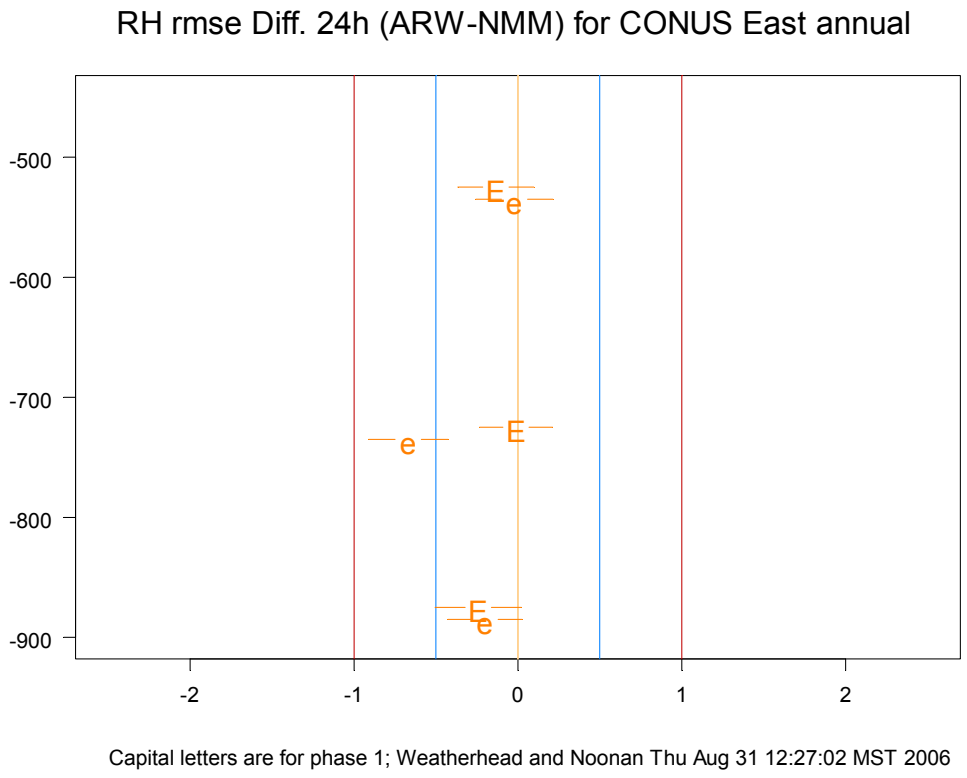
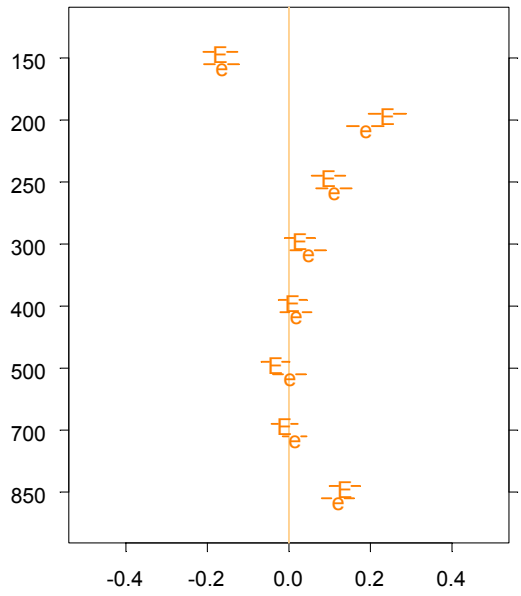


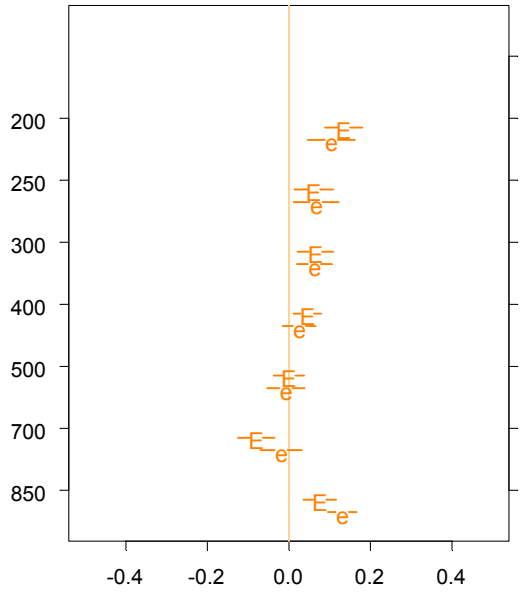
Figure C.3. Relative humidity RMS error difference (ARW-NMM) for 24-h over eastern verification. For rawinsonde verification.

Wind bias Diff. 24h (ARW-NMM) Sondes: CONUS East annual



Capital letters are for phase 1; Weatherhead and Noonan

Wind bias Diff. 24h (ARW-NMM) Aircraft: CONUS East annual



Capital letters are for phase 1; Weatherhead and Noonan

Figure C.4. Wind bias error difference (ARW-NMM) for 24-h over eastern verification. Top – rawinsonde verification, botton, aircraft verification.

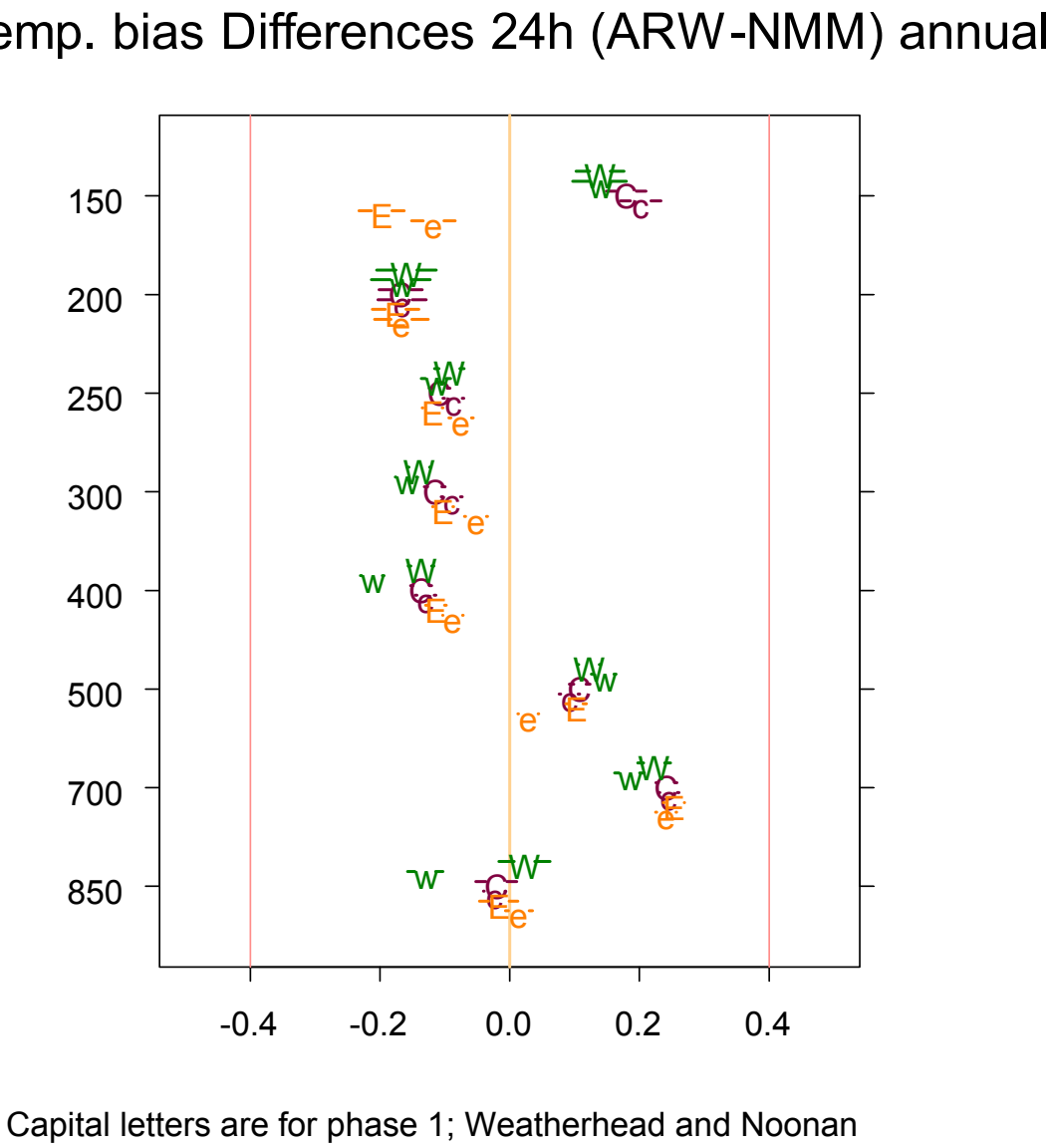


Figure C.5. Temperature bias error difference (ARW-NMM) for 24-h over CONUS, western, and eastern verification areas. Rawinsonde verification.

Appendix D – Summary comments from AWRP PDTs , DTC, and NCEP/AWC and NCEP/SPC on WRF-RR core-test evaluation

Summary (by John Brown) of PDT reports on Rapid Refresh Core Test

Turbulence (Robert Sharman and Rod Freilich, NCAR/RAL)
Full report

The Graphical Turbulence Guidance (GTG) is a statistical forecast product consisting of a mix of individual turbulence diagnostic algorithms applied to (currently) RUC, or (future) RR model output. These algorithms typically compute some measure of vertical stability, such as Richardson Number, and often derivatives of the horizontal wind field. One of the algorithms is an estimate of eddy dissipation rate (EDR) derived from second-order structure functions of the horizontal wind both parallel and perpendicular to the displacement vector. The GTG itself, and the individual algorithms that comprise it are described in Sharman et al (2006). Because GTG depends on horizontal and vertical gradients of model-predicted variables at full model resolution, the Turbulence PDT views how the models handle the scales near to the minimum resolvable as of critical importance for identifying forecastable small-scale features likely to be associated with turbulence.

The ARW and NMM outputs on native vertical levels and on the 13km RUC horizontal output grid were examined (and compared with RUC) using 2 procedures. In the first, second-order structure functions of the model horizontal velocity components and the vertical motion are computed from 6 and 12h forecasts and then compared (for the horizontal components) with those observed by Lindborg (1999) from aircraft data. All three models exhibited a more rapid falloff in the structure function than Lindborg's data show below a few hundred kilometers separation distance. For the NMM, the falloff was slightly more rapid (i.e., a slightly greater departure from Lindborg's analysis) than the ARW or RUC, which were very similar. This was interpreted by the Turbulence PDT as giving an edge to the ARW. The vertical velocity structure function with respect to horizontal displacements provided further evidence that vertical motions at smaller scales were overall weaker for the NMM than either RUC or ARW, with the difference largest for displacements less than about 8-10 Δx.

The second procedure used to examine the model outputs was to employ GTG on the ARW and NMM output in the same manner as is done operationally with the RUC. GTG was then verified against pilot reports of turbulence for both cores. Results showed a clear statistical advantage for the ARW, over both the whole verification domain (Tables 1 and 2 of the Turbulence PDT report) and, separately, the eastern domain. This advantage was noted for nearly all the individual algorithm components to the GTG as well, and the EDR in particular.

On the basis of this and the preceeding structure-function results, the turbulence PDT "strongly" recommended the ARW core.

In-Flight Icing (Ben Bernstein, NCAR/RAL)
Full report

This PDT is primarily concerned about model predictions of temperature, relative humidity, liquid and solid hydrometeors and vertical velocity. An algorithm, Forecast Icing Potential (FIP), using predictions from the RUC has been operational at AWC for a few years.

For the core test, statistical evaluation of FIP performance in predicting icing or no icing conditions as verified by pilot reports (pireps) for 60 6 and 12h forecasts initialized at 12Z showed very little difference between the cores. Likewise, distributions of temperature and relative humidity at locations of icing and no-icing pireps showed very little difference. There was, however, a tendency for stronger vertical velocities (both up and down) in the ARW at locations of pireps. Overall, considering each model grid point for both cores, the distribution of ω (= Dp/dt) is clearly broader for ARW than NMM (Fig. 5 in full PDT report). This is seen also in case examples, which show much more structure and larger magnitudes on smaller horizontal scales in the ARW as compared with the NMM. These types of features, if they could be established as meaningful, gave hope to the In-flight Icing PDT that the ARW might be capable of producing more detailed, useful icing forecasts than the NMM. For that reason, they recommended choosing the ARW.

National Ceiling and Visibility (Jennifer Simard, Paul Herzegh, Richard Bateman, NCAR/RAL)
Full report

The main concern of this PDT is the need to distinguish between flight conditions [e.g., VFR (Visual Flight Rules), IFR (Instrument Flight Rules), etc.], both current and forecast. The approach taken here was to apply the horizontal field of ceiling and visibility, diagnosed at each model grid column by the WRFpost from the 3-d model-predicted hydrometeor fields, and compare them against METAR ceiling and visibility observations. This was done separately for both phases 1 and 2, and for at least portions of each retro period, but considering only 3, 6, 9 and 12h forecasts. The metrics used for the comparisons were probability of detection (POD), False Alarm Rate (FAR), critical success index (CSI) and frequency bias for prediction of IFR conditions separately for ceiling and visibility (< 1000' for ceiling and < 3 miles for visibility). Comparison was made with output from the RUC for the same time periods.

Results of this analysis showed very similar performance by the 2 cores, and were regarded by the C&V PDT as showing no significant advantage to either

core. Both models showed some skill but tended to overforecast IFR ceilings in particular. Both were inferior to the RUC, which uses more sophisticated ceiling and visibility algorithms in its postprocessing designed to be compatible with the version of the NCAR microphysics used in the RUC.

Convective Weather (James Pinto, Cody Phillips, NCAR/RAL)

Full report

The convective Weather PDT is developing techniques for prediction of convection in the 0-6h time period. These are a blend of extrapolation procedures and forecasts based on model output, currently, RUC. In addition to the model prediction of convective precipitation itself, model-derived parameters are used including stability, moisture, vertical motion and horizontal advection of certain parameters.

The PDT evaluation consisted of examination of both model prediction of precipitation for the 6-9h period and prediction of parameters used in their automated forecast procedures. In addition several individual phase 2 cases were looked at qualitatively, and convective precipitation from the models compared with radar. It was found that both models performed acceptably and similarly in prediction of model-derived parameters. However, the tendency for the ARW to predict more precipitation amounts > 2mm/h than the NMM was regarded as a distinct advantage for the ARW in providing useful input to the blended techniques under development by this PDT. As a consequence, this PDT recommended choosing the ARW for the RR.

Comment reports by AWC and SPC below after GSD briefing via July Powerpoint summaries of preliminary results

Aviation Weather Center (Steve Silberberg)

- Previous AWC evaluation of NAM-WRF-NMM found large errors in dewpoint (Td) and RH forecasts at 925, 850, and 700 hPa. Td bullseyes of 20°C were common. The height of the marine inversion off the west coast for NMM was poorly forecast with 30°C Td errors over large areas. Also the desert SW was forecast to be too wet at 850hPa in the NMM. Water vapor analyses and forecasts remain a formidable challenge to surmount in both the NMM and ARW.
- Turbulence: Improved turbulence forecasts may be the result of ARW noise. Would go with ARW for turbulence forecasts, but not for other aviation forecasts
- Icing: Noise in ARW overwhelms the signal for vertical velocity
- ARW vertical soundings very smooth unlike the real atmosphere.
- ARW is very noisy in horizontal plots.
- ARW, like the GFS has a large positive bias for convection in air mass regimes.
- *Overall, prefers ARW for the Rapid Refresh dynamic core to provide increased diversity to forecasters as guidance.*

Storm Prediction Center (Russ Schneider, Steve Weiss)

- The test results strongly suggest that physics choices have at least as large of an impact on WRF-RR performance as the choice of dynamic core, and there are indications that model physics may play a dominant role in the first 12-24 hours of the forecast.
- Thus, SPC has no strong preference for the choice of dynamic core selected for the WRF-RR, and issues such as computer resource requirements and model run time may appropriately play a key role in determining the dynamic core used in the WRF-RR.
- The SPC operational hazardous weather forecasting mission depends most heavily on accurate real-time three dimensional analysis and short-term model forecast guidance, which requires a robust data assimilation system capable of incorporating all available observational data. It is essential that the data assimilation system for the WRF-RR replicates and preferably continues to improve upon the current RUC data assimilation system, in order to assess accurately the current state of the atmosphere and any short-term predictive changes.
- The SPC focuses much attention on the relationship between the environment and the spectrum of convective storms that can be supported by the environment. In particular, the structure and evolution of the boundary

layer is especially important, as the PBL plays a key role in determining aspects of convection initiation, intensity, and mode.

- The choice of physics for the WRF-RR, particularly the convection, microphysics, and PBL parameterizations, must be appropriately tuned and adjusted to optimize the ability of the WRF-RR to accurately analyze and predict environmental structure in a four-dimensional sense. The SPC strongly prefers a physics package similar to the current RUC physics, which tends to preserve more realistic PBL evolution compared to the current NAM physics.

The SPC is more interested in having WRF-RR model performance based on accurate analyses and short-term forecasts of the three-dimensional environment (e.g., sounding structure, derived convective parameters), with less importance placed on traditional measures of model skill such as ETS and bias scores.

DTC executive summary (31 August 2006)

Full report – DTC - Nance,
supplemental Report by Weatherhead et al.

- 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.

Appendix E: Special Experiments to Explore Cause of ARW – NMM Differences in Character of Small-Scale Features

It has been noted elsewhere (section 5.B - GSD visual evaluation report, App. D - In-Flight Icing and Turbulence PDT reports) that the ARW tends to produce more amplitude in small-scale (horizontal length scales < 10 – 15 Δx, say) features than does NMM. On the other hand, there is slightly more detail in vertical sounding structure in the NMM than in the ARW. In an attempt to better understand this paradoxical difference, suggestions were made for test runs that might make the two cores behave more similarly. This brief section discusses these experiments.

Description of experiments

One limited set of experiments was run. One-week periods were tested for both winter and summer using only Phase 2 physics (listed as "PH2r3" in the statistics shown in the WRF-RR-core-test web site at http://ruc.noaa.gov/wrf/RR/testing/NCEP_verif). The phase 2 RR core-test codes were the starting point for all code changes. These changes were as follows.

ARW – Replace the third-order upwind vertical advection used in phase 1 and phase 2 by second-order centered vertical advection. This was done easily with a change to the ARW namelist.input file. The entries v_mom_adv_order *and* v_scalar_adv_order were both changed from 3 to 2, so that vertical advection of both the momentum and scalars (such as hydrometeors) were affected. Motivation for this change was to increase vertical sounding structure by removing the damping associated with the 3rd order upwind advection as suggested by Zavisa Janjic (NCEP/EMC) and Bill Skamarock (NCAR).

NMM – Replace the horizontal divergence damping by external-wave damping using code provided by Dave Dempsey. Horizontal diffusion was left unchanged. (Tom Black –NCEP/EMC helped us to enable and test 4th order diffusion in the NMM as a possible option for this test, but visual inspection suggested that we should retain 2nd-order diffusion for the test.). Motivation for this change was to eliminate the explicit damping of internal gravity waves in the NMM by the horizontal divergence damping applied at all levels, and replace it by code that explicitly damps only the external mode.

Summary of results

Only a small sample of days (7-10 days) from the winter and summer retro periods was rerun with these changes. For both cores and for nearly every run time, verification scores were worse relative to the standard phase 2 configurations. For both cores, given that the effect of these changes was to increase the amount of small-scale detail in the forecasts, this is not surprising.

The greater small-scale detail can be seen in Fig. E1, which for each core core for the 24h forecast valid 12Z 18 January 2006, compares the phase 2 omegas with the omegas from the runs modified as described above. There is considerably more variability and amplitude in the small-scales for both cores with the respective changes. In this we see that replacement of the horizontal divergence damping in the NMM with Dempsey's external-mode damper does make the NMM vertical motion field look a bit more like the ARW's, especially away from terrain. The added small scales also seem considerably more incoherent than the larger-scale patterns that dominate with the unmodified phase-2 configurations.

ARW – Vertical soundings showed only marginal differences in mixed-layer structure for the small sample of cases available to examine. Mixed layers profiles of potential temperature and mixing ratio showed an overall tendency for more well-mixed conditions. However, this was not evident in every comparison, and the changes were only readily apparent to the practiced eye. This argues that, yes, at least part of the reason that ARW soundings tended to be smoother in the vertical was because of the 3rd order vertical advection. However, differences were subtle, and confounded by other differences in the highly convective region where the soundings were located.

Except for vertical motion, only very subtle changes in horizontal cross sections were apparent. The 850mb and the 850-500mb mean relative humidity patterns were slightly less coherent, as also the CAPE plots. There was noticeably more small-scale structure in the 500mb vorticity field. Areas of very light precipitation were increased, particularly over the oceans. Other than this, there were no systematic differences in the precipitation fields from the standard phase-2 configuration.

NMM– Along with the major increase in small-scale detail in the omega fields noted above, there was a lesser but noticeable increase in the small-scale detail in other fields. Noteworthy is that removing the horizontal divergence damping consistently increased the orographic precipitation enhancement in the West; high elevation precipitation maxima increased by ~ 10%. This is an interesting result, because the ARW consistently had greater precipitation enhancement than the NMM in the standard core-test runs. This change for phase 2 sometimes caused the NMM orographic enhancement to be a little larger than the ARW.

Conclusions

Second-order vertical advection in the ARW clearly further increased horizontal detail in the forecasts. However, one must question how much of this increased detail is meaningful, since a more accurate, but more diffusive, difference scheme was replaced by a less accurate one. Increase in detail in the vertical soundings was noticeable, and, for a small sample, did seem to improve mixed-

layer structure slightly in some cases. We think it would be worth exploring this matter further by considering additional cases and trying the fifth-order upwind vertical advection option as suggested by Joe Klemp.

Replacing the horizontal divergence damping by external-wave damping had a dramatic effect on the small scales of the vertical motion field of the NMM, particularly away from mountains. We might have seen a larger effect over the West had we looked at vertical motion at high levels where vertically propagating waves can be of large amplitude. Eliminating the divergence damping in NMM T-REX runs increased wave reflection from the top boundary after 12h or so. Aside from the possible benefits of enhancement of orographic precipitation, it is not evident to us that eliminating the horizontal divergence damping in the NMM would have any advantages from a strictly forecast perspective.

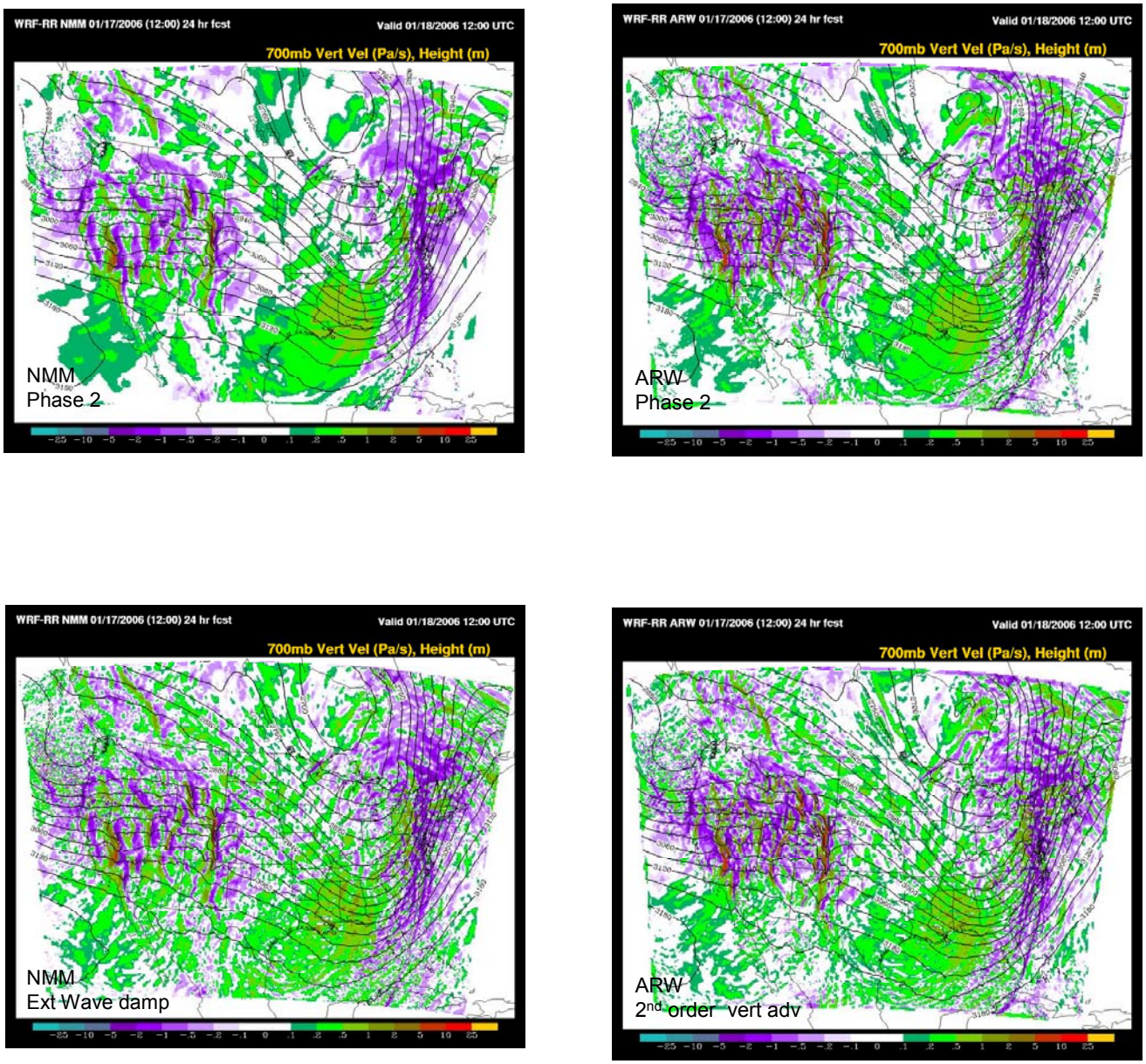


Figure E1. Four 24-h forecasts of 700mb omega (= Dp/Dt, pa/s), all valid 12Z 18 January 2006. The top two are the phase 2 results for the NMM (left) and ARW. The bottom 2 are using the modified phase-2 code as described in the text, NMM (left) and ARW.