22nd Conf. Wea. Analysis Forecasting / 18th Conf. Num Wea. Pred., June 2007, Park City, UT, Amer. Meteor. Soc.

4A.1

Rapid Refresh Core Test: Aspects of WRF-NMM and WRF-ARW Forecast Performance Relevant to the Rapid Refresh Application

John M. Brown[1], Stan Benjamin[1], Tanya Smirnova[1,2], Georg Grell[1,2], Ligia Bernardet[1,3,4], Louisa B. Nance[5], Randall Collander[1,3], and Chrstopher W. Harrop[1,2]

[1] NOAA Earth System Research Laboratory, Boulder, Colorado
[2] Collaboration with the Cooperative Institute for Research in Environmental Sciences (CIRES), Boulder, CO
[3] Collaboration with the Cooperative Institute for Research in the Atmosphere (CIRA), Fort Collins, CO
[4] Also affiliated with NCAR-NOAA Developmental Testbed Center
[5] NCAR-NOAA Developmental Testbed Center and NCAR/MMM

1. What is the Rapid Refresh?

In the year 1994, the National Centers for Environmental Prediction (NCEP) introduced the first US operational model to be run at intervals less than the standard 12h between rawinsonde observations. Dubbed the Rapid Update Cycle, or RUC, this model was initially run as the forecast component of an intermittent (3-h interval) assimilation cycle over a domain slightly larger than the coterminous US. This more rapid updating was made possible by increasing number of “offtime” observations, particularly from commercial aircraft and from the experimental NOAA Profiler Network that had come into existence a few years earlier. Over the years, both the analysis and model have undergone substantial upgrades and this model is now run in a configuration with 13km horizontal grid spacing and 50 hybrid-sigma-isentropic levels over a domain covering the coterminous US, southern Canada, northern Mexico and adjacent waters at an assimilation frequency of one hour. A full description of the RUC as it existed in 2003 can be found in Benjamin et al (2004a,b). A discussion of upgrades since that time can be found in Benjamin et al (2006,2007).

The RUC occupies the “situational awareness” niche in the NCEP model suite. That is, forecasters use it extensively as an aid in monitoring the latest trends in fast-breaking weather situations for the purpose of updating very short-range forecasts. The primary users of the RUC are therefore not surprisingly forecasters concerned with severe local storms and with weather having a high impact on aviation, both from considerations of safety (e.g., turbulence) and operational efficiency (e.g., flight routing).

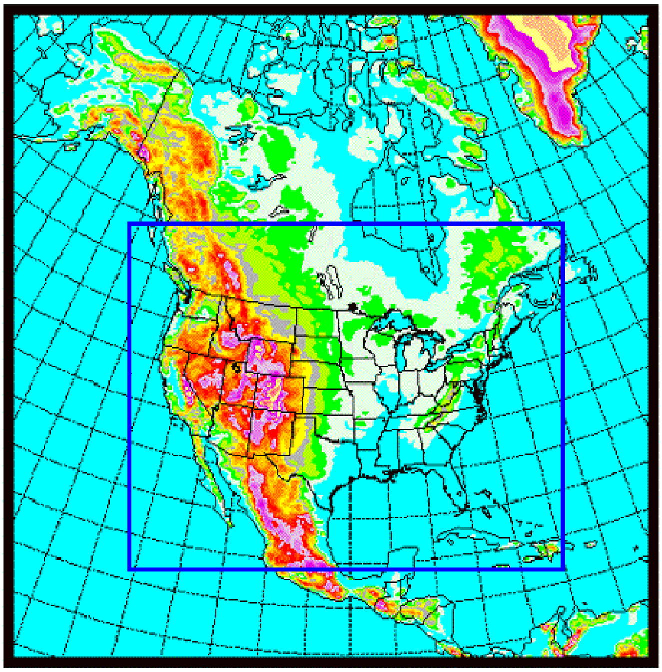


Figure 1. Approximate Rapid Refresh domain (outer region). Blue rectangle is the current RUC domain.

Looking ahead to the time when it would be necessary to use a non-hydrostatic model as the forecast component of the RUC, and toward a NOAA goal to accelerate transition of development from the research community into operational environmental models, NOAA/NCEP and NOAA Earth System Research Laboratory (ESRL) Global Systems Division (GSD) (formerly NOAA FSL) agreed in 2002 to select a version of the Weather Research and Forecast (WRF) model to replace the current hydrostatic Rapid Update Cycle (RUC) model. A year later, in response to desire of the National Weather Service (NWS) to make forecasts for Alaska as part of the RUC, it was decided to more than double the size of the current RUC domain to include most of Alaska, as well as Puerto Rico and the US Virgin Islands, while keeping the rapid update function intact. This new analysis and nonhydrostatic model forecast configuration was dubbed the *Rapid Refresh (RR)*. The RR domain as it is currently configured in testing at ESRL/GSD is shown in Fig. 1, along with the present RUC domain.

The WRF model has a number of options, including two primary versions of its dynamical core: the Advanced Research WRF (ARW, Skamarock et al 2005), and Non-hydrostatic Mesoscale Model (NMM, Janjic and Gerrity 2001) cores. By the term “dynamical core” we refer to that part of the model code necessary to run the model adiabatically, i.e., without any physical parameterizations and without diffusion except that either implicitly or explicitly included in the finite-difference approximations to the model governing equations for the sole purpose of controlling the buildup of energy in the smallest resolvable scales due to nonlinear advective processes.

In accord with the 2002 agreement between NCEP and ESRL, NOAA/ESRL’s Global Systems Division in collaboration with the NOAA/NCAR Developmental Testbed Center (DTC) and the National Center conducted a 9-month objective evaluation for Atmospheric Research (NCAR). This testing of the ARW and NMM dynamic cores was unique relative to 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 dynamical core for the specific aviation and severe weather applications planned for the Rapid Refresh* at the expected initial Rapid Refresh horizontal resolution (~13km).

In this paper, we describe the outcome of this evaluation. But, first it is necessary to set the stage by summarizing the experimental design, RR-specific evaluation criteria and modifications to the ARW and NMM cores necessary to isolate differences in performance due to the dynamical cores rather than other factors, for example, slightly different versions of physics schemes running in each core. A companion paper by Nance et al (2007) contains a more complete description of the statistical evaluation and certain aspects of the experimental design.

2. Ground rules for the evaluation

A set of ground rules for how this comparative evaluation of the NMM and ARW core would contribute toward the decision of which core would be used for the RR was established by NOAA/ESRL/GSD in a 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. Experimental design

The overall goal of the RR Core Test was to conduct experiments as broad in scope as possible, given resources available, and to ensure controlled experiments, as much as possible. The highest priority was to ensure that the experiments were controlled, that is that the model configurations, including physical parameterizations, were as identical as possible without having to make changes that would entail wholesale alteration of the cores themselves. This entailed various trade offs. In the following subsections we consider all aspects of the experimental design. More details are in Nance et al. (2007). The WRF-RR core-test project 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 practical to that we anticipate for the RR when it becomes operational. However, with limitations of computer power and verification capability, we recognized early on that we were not going to be able to run over the full RR domain of Fig. 1. We also regarded it as important to do repeat runs with 2 different physics configurations or suites in order to reduce the possibility that considering just one physics configuration might favor one core or the other. We desired to have one of these physics configurations be close to what we currently use in the RUC (so-called “RUC look-alike” physics), given that we intend to use in the RR at least some of the physics packages (or updated versions of them) currently in the RUC.
Other aspects of the current Rapid Update Cycle, including ~13-km resolution (close to what we anticipate for the initial operational RR implementation), emphasis on short-range forecasts, and use of RUC initial conditions including 3-d hydrometeor fields, were also desired. The intention was to conduct experiments as broad in scope as possible, given resources available, and to ensure controlled experiments, as much as possible. The experiment preparation procedures are broken down into model code modification and experimental design.

a. Code modifications to WRF model necessary for evaluation

Previous informal attempts to compare the ARW and NMM cores have been seriously limited by a number of factors. A major one has been the choice of different physics packages when the 2 cores have been run using the same initial data (Kain et al 2006). Therefore, a requirement for the RR core Test was to use identical physics suites in each core, including specifics of feedbacks between individual physics schemes and with the dynamics. We also set a goal to do repeat runs with the same initial and boundary conditions, but with 2 different physics configurations or suites in order to reduce the possibility that considering just one physics configuration might favor one core or the other. We desired to have one of these physics configurations be close to what we currently use in the RUC (so-called

"RUC-like" physics), given that we intend to use in the RR at least some of the physics packages (or updated versions of them) currently in the RUC. Since at the time our experiments were being designed the WRF-NAM (implemented operationally in June 2006) was under test with very different convective parameterization and microphysics schemes than are used in the RUC, it was decided to use the WRF-NAM physics suite for the other configuration. Since this suite was closer to interoperability in both cores than a "RUC-like" suite, the first set of core-test runs was with this suite. We refer to the core-test runs with the spring 2006 NAM physics as "phase 1" of the core test. The set of runs (using same initial and boundary conditions for the same initial times and same domain as phase 1), but using the "RUC-like" physics we call phase 2. These physics suites are summarized in Table 1.

Parameterization	**Phase 1 – NCEP physics suite**	**Phase 2 – RUC physics suite**
Explicit clouds	Ferrier (2006, personal communication)	Thompson-NCAR (Thompson et al 2004, Thompson 2006 personal communication)
Sub-grid-scale convection	Betts-Miller-Janjic (Janjic 1994, Janjic 2006 personal communication)	Grell-Devenyi (Grell and Devenyi 2003)
Land-surface	NAM/F77 version of Noah ("99" LSM) (Chen et al 1996, Ek 2006 personal communication)	RUC-Smirnova (Smirnova et al 2000)
Turbulence mixing	Mellor-Yamada-Janjic Janjic 1994, 2001)	Mellor-Yamada-Janjic
Radiation	Longwave/Shortwave – GFDL	Longwave /Shortwave – GFDL

Table 1. Physical parameterization suites used in the WRF-RR core-test evaluation. Additional references can be found in Skamarock et al (2005).

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 of the evaluation.

Parameterization	**NMM**	**ARW**
Ferrier microphysics	✓	Radiation problem
NCAR-Thompson	-	Old version in WRFv2.1

microphysics		
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.

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. Enhanced physics or initial-condition availability resulting from the WRF-RR core-test project. Bold, blue checkmarks indicate new or modified availability from the core-test project.

These modifications were made available to the WRF Repository, and most were included in the WRF v2.2 release in December 2006, significantly enhancing physics interoperability for the whole worldwide WRF community.

b. Domain, initial and lateral-boundary conditions

A forecast model has to be sufficiently robust to run reliability in all meteorological situations that might be encountered over its domain. This requirement dictated that RR Core Test runs be conducted over all seasons, not just one or 2, in order to compare core performance over a wide variety of weather situations. We also kept the duration of runs short, since the RR forecast duration is unlikely to be more than 12-18h. (Keeping the run duration short also allowed us to make more forecasts.) To accomplish this, we did the following for each phase:

- 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
- On these days, runs were initialized at 0000 and 1200 UTC, and run to 24h.

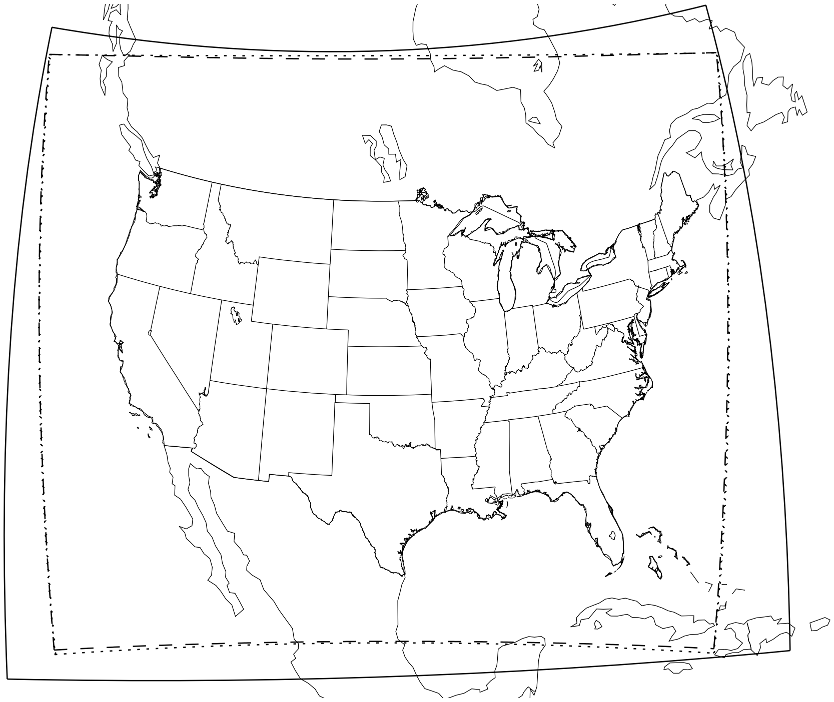


Figure 2. ARW and NMM forecast domains for WRF-RR core test. Solid line is RUC13 451x337 domain. Dashed and dotted lines are for ARW (Lambert conformal) and NMM (rotated lat/lon) domains.

With limitations of computer resources and verification capability, we recognized early on that we were not going to be able to run over the full RR domain of Fig. 1 at a resolution approaching that planned for the operational RR. We also wished to initialize with RUC initial conditions so that we could minimize spinup of clouds and precipitation during the early hours of the run. (In the RUC we cycle the 1-h forecast of hydrometeors as well as of other quantities as background for the next hour's analysis, including analysis of hydrometeors—see Weygandt et al. 2007), and plan to do the same in the RR. These 2 constraints led to the decision for constructing nearly matching sub-domains of RUC CONUS domain [Fig. 2, more detail in Nance et al. (2007)]. Some small differences were unavoidable since no single map projection is available for both cores.

RUC13 native coordinate data from GSD dev13 RUC runs for the atmosphere and land surface, including hydrometeors and the NCEP SST-0.5 deg daily updated analysis enhanced by special treatment for lakes, were used to initialize.

Lateral boundary conditions were from the NAM AWIPS 212 grid (40km) using the NAM run initialized 6h previous to current WRF initial time, as is done for the boundary conditions of the operational RUC.

c. Model configuration

1) MODEL PHYSICS

Model physics for Phases 1 and 2 were as shown in Table 1.

2) HORIZONTAL GRID SPACING

Horizontal grid spacing was approximately 13km as discussed in more detail in Nance et al (2007).

3) TIME STEPS

It was ultimately decided to choose the time steps as close to the maximum stable values for each core, as determined by consultation with developers and our own experimentation, rather than trying to make them such that the interval between physics calls was identical between the 2 cores. A factor in this decision was the current inability to call microphysics less frequently than every large time step in the ARW.

- ARW – Large time step: 72 s, small (acoustic) time step: 18s; physical parameterizations (except radiation) called every long time step.
- NMM – 30 s for dynamics, 60 s (2 dynamic time steps) for physical parameterizations other than radiation as well as advection of water vapor and hydrometeors.

- These time step lengths were shown to be the maximum stable values in testing for the other configurations.

4) VERTICAL LAYERS

The number of vertical levels was set to 50 (i.e, 49 layers) for both cores, with their vertical distribution set similarly for ARW and NMM. (The vertical distribution 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.)

5) MODEL TOP

Top of model domain was restricted to be no higher than 50 mb due to use of RUC initial conditions, since top of RUC model is ~ 50mb. (This top will be raised for the operational Rapid Refresh).

6) PROVISION FOR WAVE REFLECTION FROM TOP BOUNDARY

Upper boundary condition for both cores is zero vertical velocity. This boundary condition is reflective of vertically propagating waves. In addition to damping of gravity-wave energy implicit in the numerics of the two cores, the ARW provides for the possibility of introducing an upper-level wave-damping layer through the model namelist (See Benjamin and Brown 2006, Appendices B1 and B2). In the core test, this was selected (after limited experimentation) to occupy the top 5km of the model domain with a damping coefficient (see Skamarock et al 2005 for details) of 0.02. There is no provision for such upper-level damping in the NMM.

7) TERRAIN ELEVATION SPECIFICATION

Due to differences in the SI software for vertically interpolating RUC native-coordinate data to the NAM and ARW vertical coordinates, there are some differences 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 in Fig. 3. Therefore, we have generally preferred verification results from the Eastern US verification domain vs. the Western US domain or the full (“National”) verification domain.

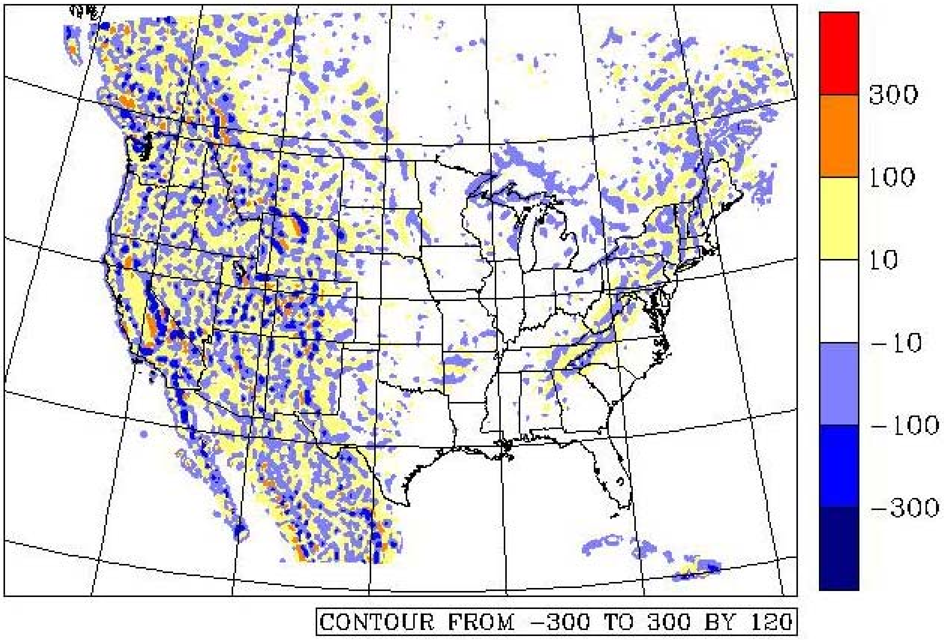


Figure 3. Terrain difference – ARW value minus NMM value. (Nance)

d. Post-processing

The WRF post-processing routine (*WRFpost*, largely developed at NCEP) was used to produce isobaric grids and other diagnostic products from the mass-coordinates of the native grids for ARW and NMM models (see Nance et al, 2007, 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 by means of 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 and also over two regional sub-areas (Fig. 4). Additional detail on this is provided in Nance et al (2007).

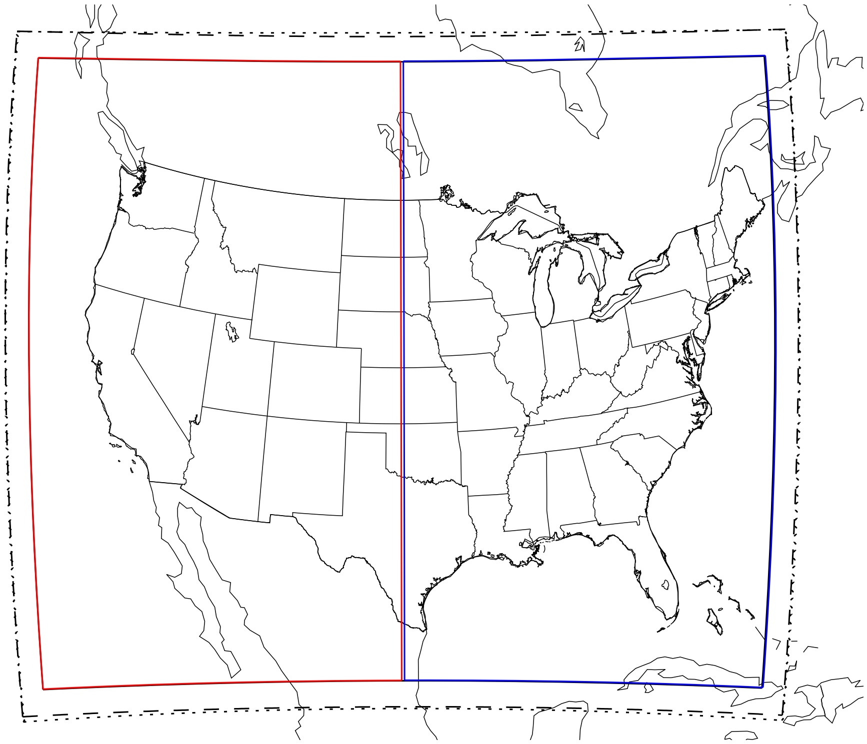


Figure 4. Verification areas, over a) full CONUS area (just inside full model domain) (on left) and b) western and eastern regional verification areas (on right).

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. This was indeed the case, lending support to the verification results discussed here.

4. Background considerations and evaluation criteria

We provide here background information to set the context of the evaluation, and criteria by which the model forecasts were evaluated. As noted previously, these criteria were intended to focus the evaluation on aspects of model performance we considered particularly relevant to the anticipated heavy use of the Rapid Refresh as the primary situation awareness and very short-term forecast model run operationally at NCEP.

a. Strongly constrained experiments

The experimental design was strongly constrained, more so than other model comparison studies done in the past at NCEP. Specifically:

- 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, based on our experience with RUC, 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 chosen with these constraints in mind.

b. Verification against various observation types

1) 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 too low an elevation angle for accurate tracking of balloons 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. Further, 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).

2) 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.

3) PROFILER

Profiler observations were not used due to an error in the NCEP observation processing whereby height-to-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 evaluated for their statistical 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 made by Weatherhead et al (2006a,b) and Weatherhead and Noonan (2006a,b). Their techniques are described in detail in Weatherhead et al (2006a,b). GSD also performed an initial statistical significance comparison using the Student's t-test score. Only Weatherhead et al results are shown here, 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;
- Accounts for auto-correlation of differences in estimating degrees of freedom for significance;
- Separate statistics produced for
    - 3 verification areas (Fig. 4): CONUS, western, and eastern,
    - Phase 1 and 2 physical parameterization suites,
    - All four seasons are considered separately and jointly.

2) SUBJECTIVE SIGNIFICANCE

As part of this project, prior to conducting the core-test runs, GSD established subjective evaluation criteria for *magnitude of error difference* averaged over 1-month periods (Table 4) 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 Table 4 criteria 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:
1) Differences should be consistent over the majority of individual verification times over a month-long seasonal period.
2) 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 °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 5a, but less heavily than RMS error.

In the GSD analysis of upper-air verification (section **5),** vertical lines are plotted 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 the surface verification 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 the “subjective” and “statistical” criteria defined in 1) and 2) above *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.

The 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 than the western area, 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. 3).

5. Evaluation of Core-Test_experiments by ESRL/GSD

Using the general considerations discussed and procedures outlined in Section 4, and some more specific ones listed below, the results of the evaluation by GSD are summarized in this section. All the normalized difference plots in Section 5a except Fig. 5 are taken from an exhaustive report by Weatherhead and collaborators (Weatherhead et al 2006b). Formal statistical significance at the 95% level is identifiable in these plots 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.** Conversely, almost all model difference in RMS errors showing formal statistical significance also met subjective significance criteria ("yellow zone" – concern).

a Statistical evaluation

The following summarizes factors we considered in the evaluation to follow.
1. We regard the mandatory isobaric levels considered here (850, 700, 500, 300, 250, 200mb) as all approximately equal in importance for this test. Jet aircraft generally cruise in the 300 – 200mb layer, so accurate winds at these altitudes are needed for fuel efficient flight routing. However, winds on descent are crucial for precise arrival coordination at major airports. 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.

2. Results for rawinsonde and aircraft verification should corroborate each other, subject to the error and distribution differences described in section 4b.

3. "Significant seasonal differences" (SSDs**)** for a given variable and level and physics suite were defined in section 4c as month-long statistical differences between ARW and NMM cores for 24-h forecasts verified over the eastern verification region (Fig. 4) considered significant from both formal statistical and subjective perspectives.

1) 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. 5 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. 5a to determine which model was more accurate, on the average. For instance, the ARW model with Phase 1 (Ph1) physics (marked with a blue **x**) can be compared with the corresponding NMM model also with Ph1 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.

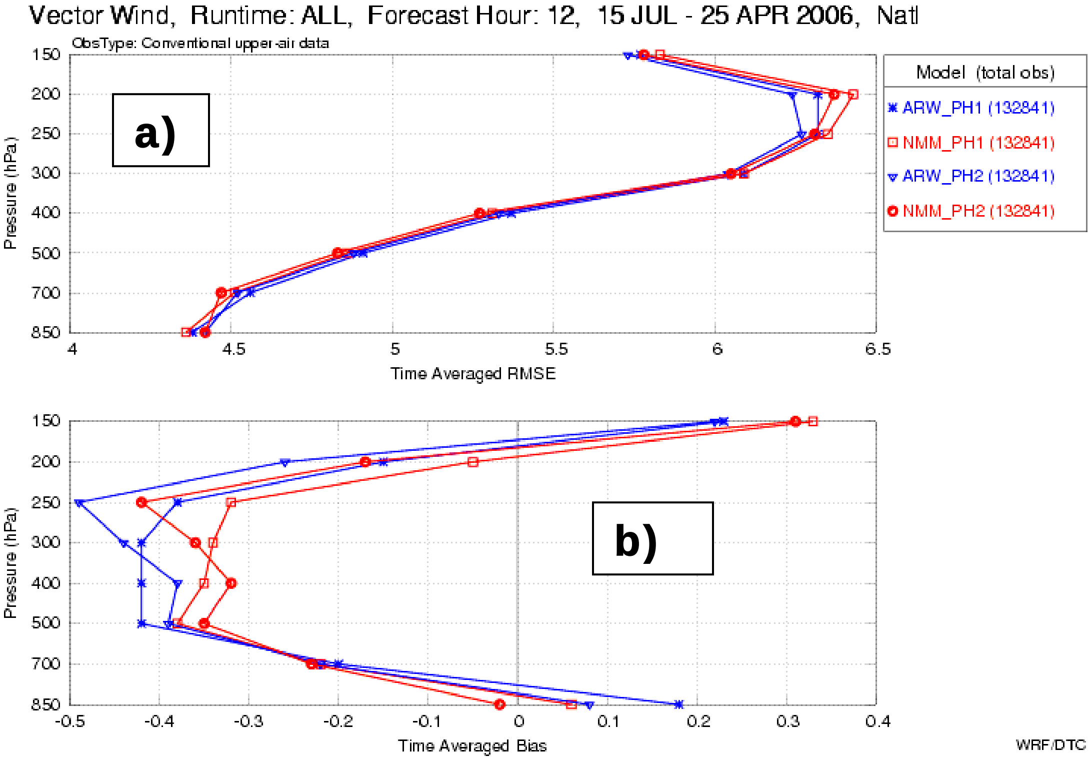
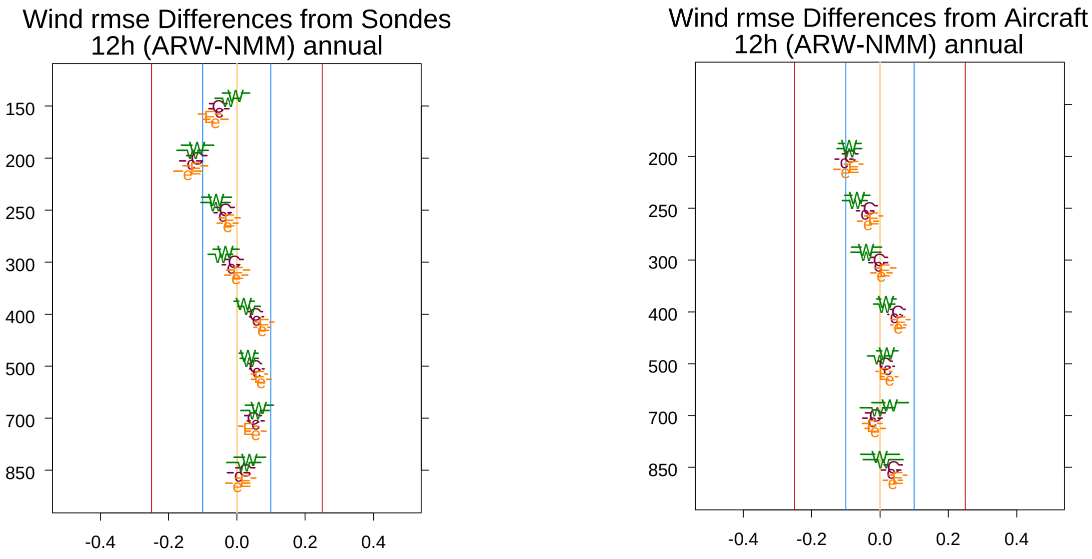


Figure 5. 12-h RMS vector wind error (a - top) and bias (b- bottom) vs. rawinsonde observations averaged over all 4 seasons for 4 different WRF model versions, ARW and NMM cores, each with Phase 1 and Phase 2 physics. Units in m/s. (From WRF-DTC verification display at http://ruc.noaa.gov/wrf/RR/testing/NCEP_verif .)

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 (2006a). 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 4c) indicating significant differences in errors between different model versions.

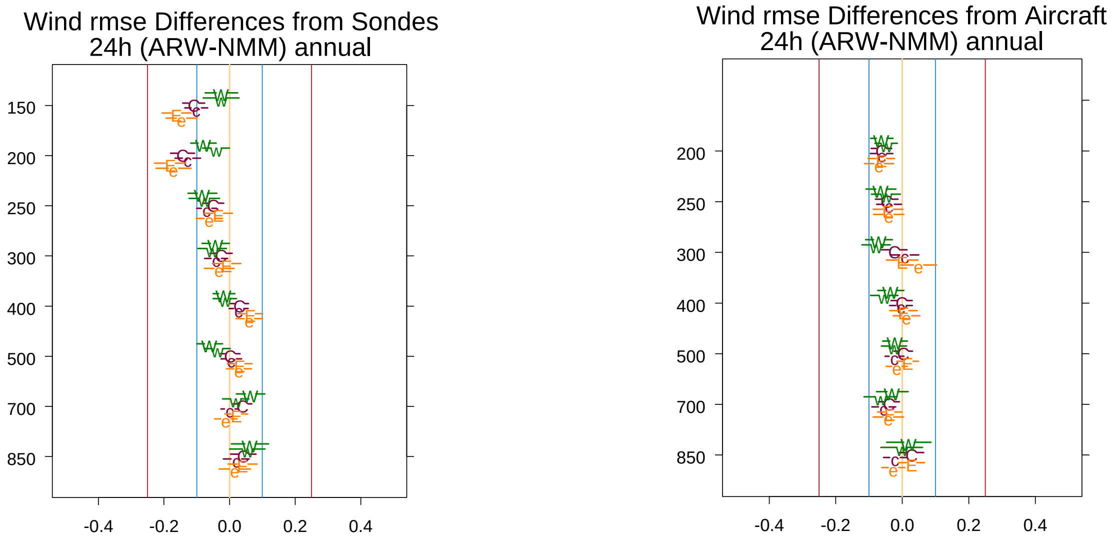
In Fig. 6, wind verification results show more accurate 12-h forecasts from the ARW at upper levels and, to a lesser extent, from the NMM in the lower troposphere, averaged over the 4-season test periods. These differences are muted when verified against aircraft reports: smaller advantage for the ARW at upper levels, and almost no advantage for the NMM in the lower troposphere. For 24-h forecasts (Fig. 7), 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. 7 in Fig. 8 showing only the eastern area. The justification for using the eastern verification area (section 4c) is evident in Figs. 7 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. 9) and aircraft verification (Fig. 10). 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).


Wind rmse Differences from Sondes
12h (ARW-NMM) annual

Wind rmse Differences from Aircraft
12h (ARW-NMM) annual

Figure 6. 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.


Wind rmse Differences from Sondes
24h (ARW-NMM) annual

Wind rmse Differences from Aircraft
24h (ARW-NMM) annual

Figure 7. Same as Fig. 6 but for 24-h forecasts.

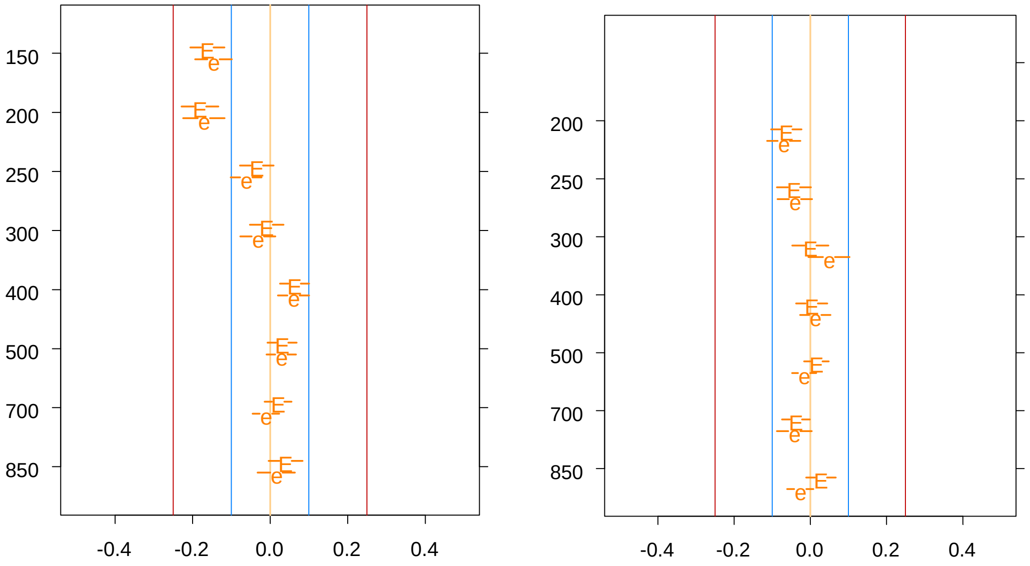


Figure 8. Same as Fig. 7 (ARW-NMM difference in RMS wind vector for 24-h forecasts, rawinsonde on left, aircraft of right), but now showing *eastern verification area statistics* only. Large case 'E' is for Phase 1 physics, and small case 'e' is for Phase 2 physics.

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 4c) were determined. 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. 9), 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 SSDs (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 used for verification above 200hPa), leading to 56 possible SSDs per variable. When the same evaluation is completed with aircraft observations (Fig. 10), 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**, due to better representation of high windspeed situations with aircraft data (see discussion on this in section 4b). 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 absolute value than the ARW model (Fig. 11), 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 4c) 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 Weatherhead and Noonan, 2006a).

2) 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. 11. 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 1000-850 hPa aircraft wind bias, we consider this bias to reflect a positive speed bias at the lowest few computational levels of each core, with the ARW bias larger.

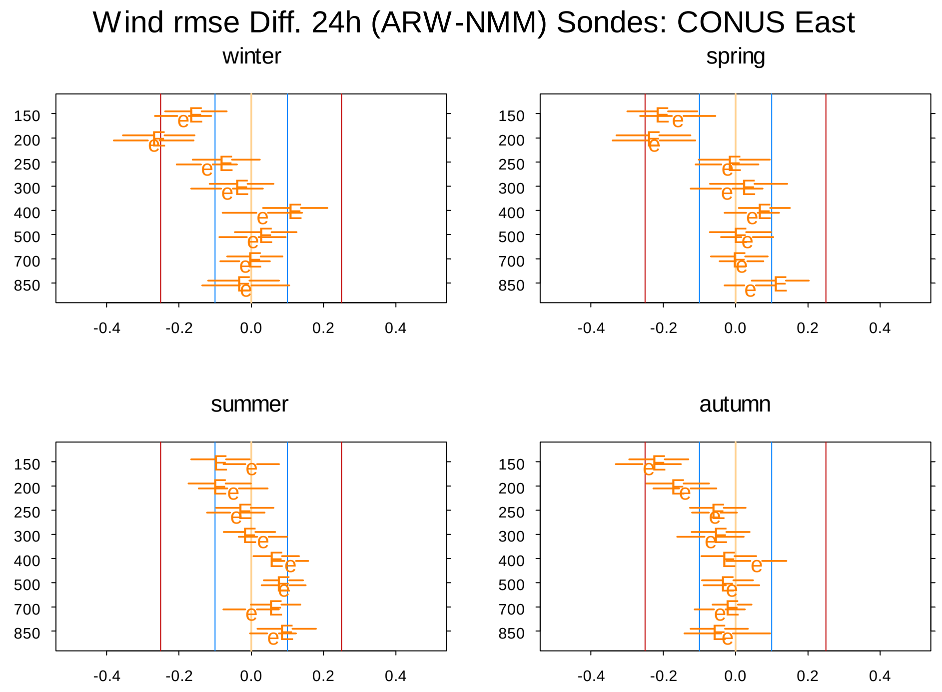


Figure 9. Wind RMS vector error ARW-NMM differences (units in m/s) verified with rawinsonde observations for 24h forecasts (eastern verification region only), but now broken down by each of the 4 month-long season periods. Blue and red lines are shown at "concern" (± 0.1 m/s) and "serious concern" (± 0.25 m/s) differences, as described in section 4.c.3. Large case 'E' is for Phase 1 physics, small-case 'e' is for Phase 2 physics.

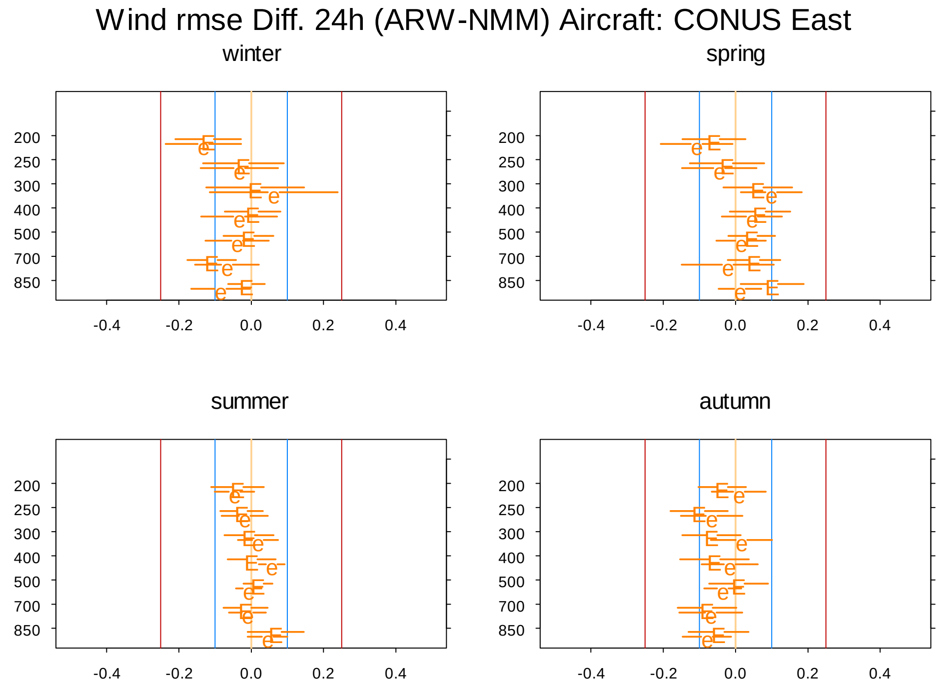


Figure 10. Same as Fig. 9 but verified with aircraft observations

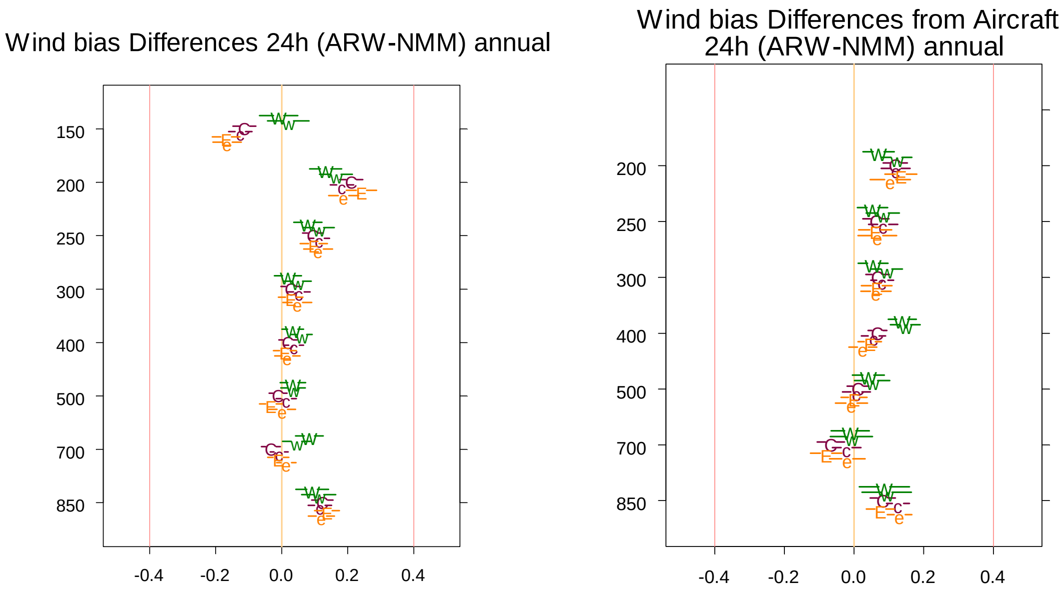


Figure 11. Wind speed bias difference for verification against a) rawinsonde (left) and b) aircraft (right). Absolute values are applied first to both ARW and NMM wind speed bias at each verification time. Positive number means that ARW absolute wind bias is larger than that from the NMM.

3) TEMPERATURE ABOVE SURFACE

Temperature 24-h forecast errors are shown in Fig. 12 verified against rawinsonde observations and in Fig. 13 against 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 4c), 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 only for Phase 1. Phase 2 forecasts (not shown) are 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. One SSD favored the NMM (850-700 hPa), with a total score favoring ARW – 3-1.

We judge the RMS temperature forecasts as similar overall but slightly favoring the ARW core in the lower troposphere. Annual average temperature error differences (not shown) also indicate smaller ARW errors in this same layer.

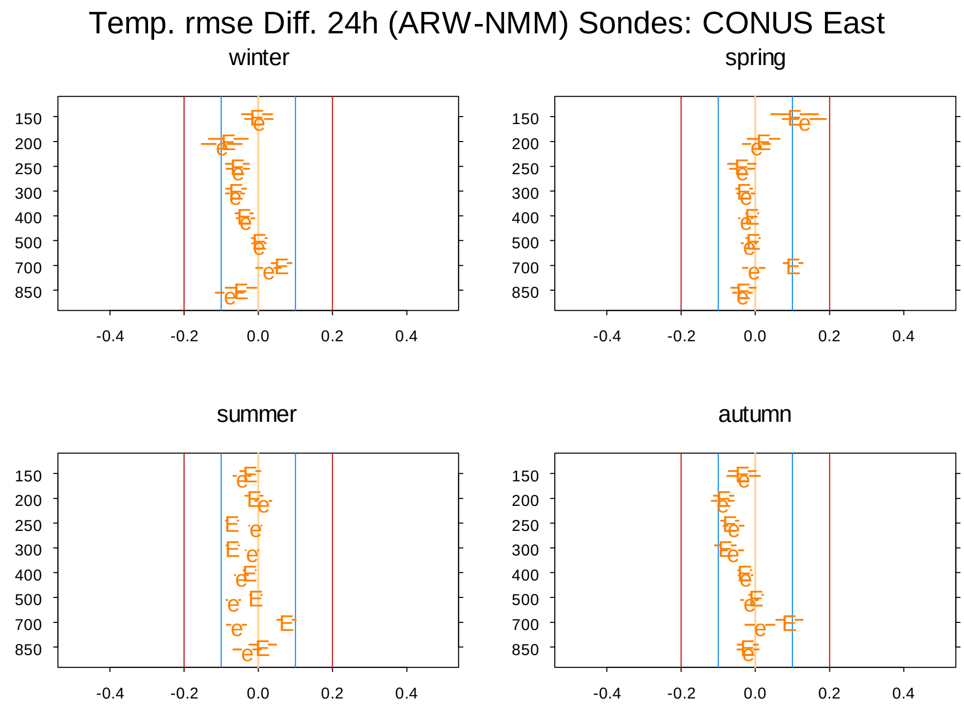


Figure 12. Temperature RMS error ARW-NMM differences (units – deg C) verified with rawinsonde observations for 24-h forecasts, broken down by each of the 4 month-long season periods. Blue lines are shown at ± 0.1 °C, corresponding to the "yellow zone" subjective significance criteria described in section 4c. Red lines are shown at ±0.2 °C, corresponding to "red zone" criteria. Only eastern verification area statistics are included.

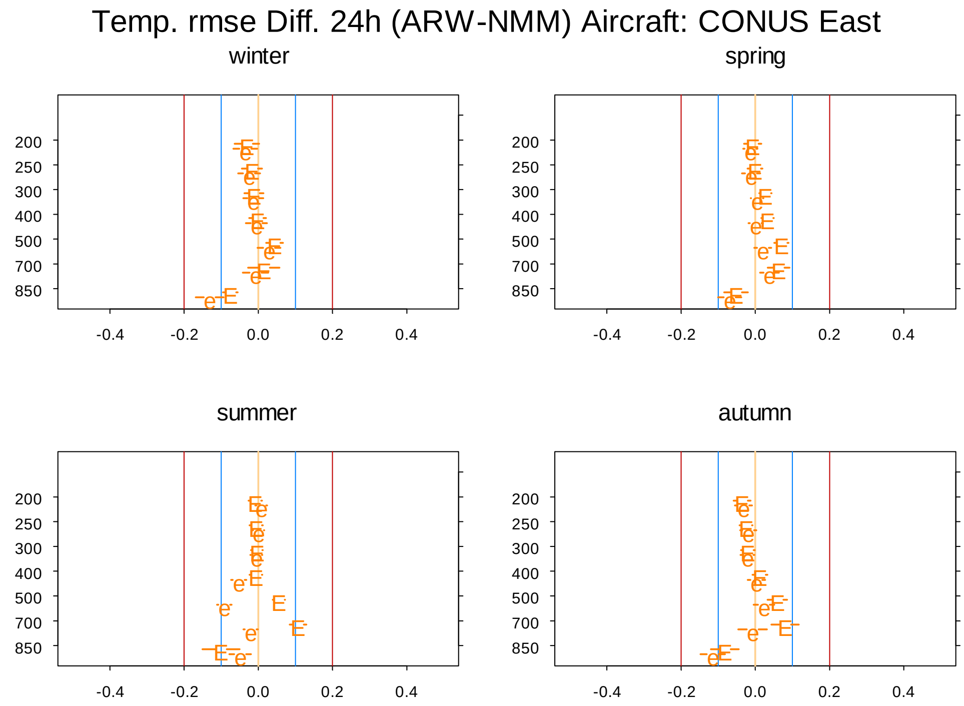


Figure 13. Same as Fig. 12, but verified with aircraft observations.

4) 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 4b, this field is dependent on reduction to 2-m level and we consider it suspect for meaningful verification.

5) RELATIVE HUMIDITY

As shown in Fig. 14, for relative humidity above the surface, the ARW gave lower error in colder seasons (winter, fall) at 850 hPa, whereas the NMM gave lower errors in summer. Only 3 mandatory levels were used for RH (see section 4c), 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).

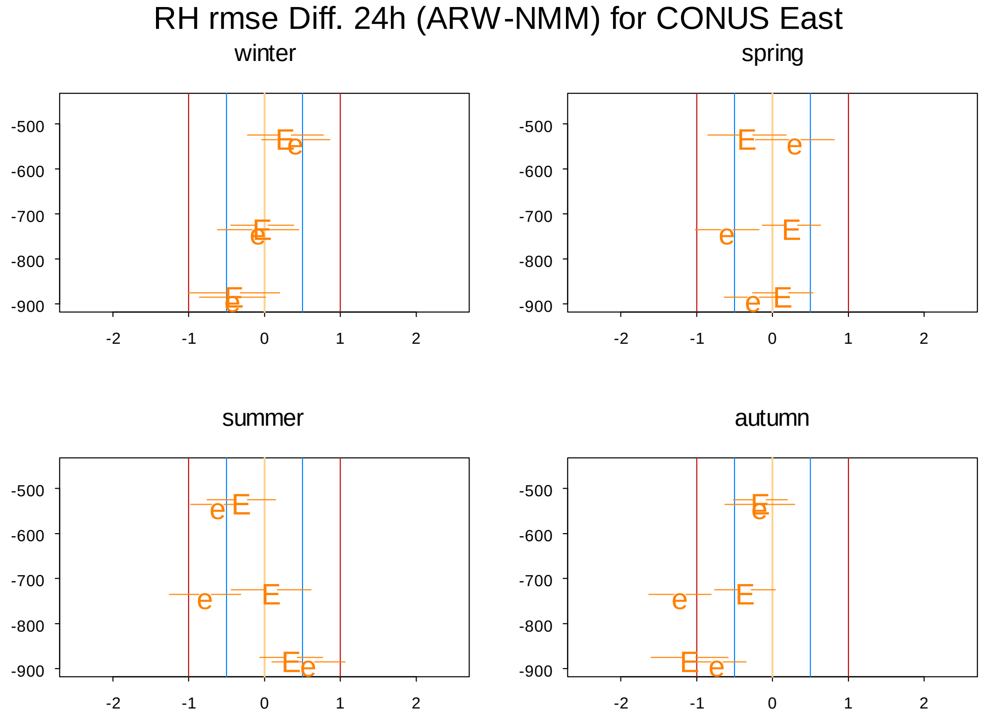


Figure 14. Relative humidity RMS error ARW-NMM differences (units – %RH, 0-100) verified with rawinsonde observations for 24-h forecasts, broken down by each of the 4 month-long season periods. Blue lines are shown at ± 0.5 %RH, corresponding to the “yellow zone” subjective significance criteria described in section 4.C.2. Red lines are 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 (not shown) indicate 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.

6) PRECIPITATION

A summary of the total 4-season precipitation forecast skill [equitable threat score (ETS) and bias] is shown in Fig. 15. 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.

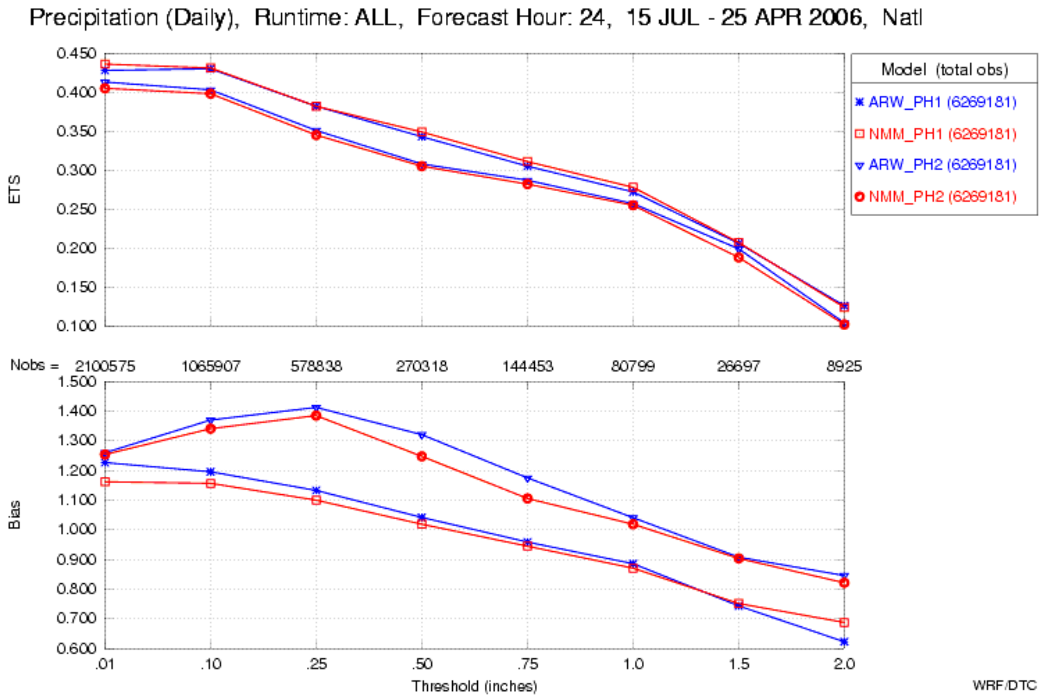


Figure 15. 24-h precipitation verification over all 4 seasons for 4 WRF model versions: ARW-Phase 1 physics, NMM-Phase 1, ARW-Phase 2, and NMM-Phase 2. Equitable threat score (ETS) on top and bias on bottom. (From WRF-DTC verification display at http://ruc.noaa.gov/wrf/RR/testing/NCEP_verif .)

6. Discussion and conclusions

This paper and a companion to be presented by Nance et al (2007) report on what, to date, is the most rigorous forecast comparison yet conducted of the two WRF cores currently available to the general WRF community, the ARW and NMM. This paper has concentrated on aspects of performance of particular concern to the Global Systems Division of the Earth System Research Lab of NOAA, namely the choice of core that will best serve as the forecast model in the NCEP Rapid Refresh. Because of anticipated heavy use of the RR by aviation forecasters and, in the future, as primary input into a four-dimensional database of probability of encounter with weather conditions impacting flight safety and flight operations, this ESRL/GSD report gives some emphasis to these aspects of performance. To aid us in this aspect of the evaluation, we solicited input from Research Teams (formerly Product Development Teams) of the Aviation Weather Research Program of the Federal Aviation Administration.

It should be emphasized that these results cannot be necessarily generalized to performance of these cores at distinctly smaller or larger horizontal grid resolutions or domain sizes. The combination of domain size and grid spacing of this particular core test permit resolution only of phenomena that are sufficiently large that the hydrostatic approximation should be closely satisfied. This is the case even though the dynamics of both WRF dynamic cores allow for nonhydrostatic motions. This means that these results are not *a priori* generalizable to horizontally smaller nonhydrostatic scales, in particular, convection-resolving forecasting such as described by Kain et al (2006). Likewise, of course, we cannot generalize our conclusions to cover global-scale processes or those occurring in the equatorial regions.

a. Relative strengths of either core

1) ARW

Major advantages

- Upper-level wind. This is apparent in aircraft verification. Rawinsonde verification (where ARW advantage was even stronger) is considered flawed [see section 5a1)].
- Lower-troposphere temperature [see section 5a3)]
- Lower-troposphere relative humidity, primarily at 850 hPa, considered to be potentially important for icing and ceiling forecasts. [See section 5b5)]

Secondary advantages

- Community involvement – Currently much more significant with ARW testing and applications than with NMM. This may change as NMM receives additional 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

Major advantages

- Wind speed bias, particularly at upper levels [section 5a1)].
- Precipitation bias [section 5a6)]

Secondary advantages

- Code already developed for 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 IGNORED OR REQUIRING FURTHER INVESTIGATION

These are matters of difference which we considered were either irrelevant or about which more investigation will be required.

1. Execution speed: faster speed will allow higher spatial resolution, assuming that the “effective resolution” is also equal. We were not able to do any timing tests on an IBM computer similar to NCEP’s, or to do timing tests over the RR domain of Fig. 1. Tests on GSD’s Ijet computer performed by Tanya Smirnova suggests that the execution speed of the NMM is roughly equal to or faster (by 10 – 40%) than ARW when physics is called at the same interval in both cores, depending on physics suite used, and also taking into account that a new ijk index-order version of NMM is 10-15% faster than the NMM version used in this core test. A 40% difference in execution speed is equivalent to only about a 12% difference in horizontal grid spacing.

2. Changes to improve performance to both cores and physics suites have been made since the configuration used in this core test was decided upon. Models are a moving target, and the results of this core test may not hold with current or future configurations of either core.

3. We considered the difference in amount of small-scale detail in forecast output noted in Section 5b as not an inherent advantage for either core. It was not a factor in our recommendation.

4. 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.

b. ESRL/GSD recommendation

- Based on the evaluation reported here, 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 2009.
- 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.

- Our recommendation is for the *initial* operational implementation of Rapid Refresh currently planned for 2009. 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.

- 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)
    - Research Teams for aviation weather funded through the FAA Aviation Weather Research Program

- 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.

7. Acknowledgments

The WRF-Rapid Refresh core test evaluation has been dependent on a wide group of participants from GSD, DTC, and NCAR beyond the authors of this paper. 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. Barry Schwartz (Schwartz Weather Consultants) performed verification and significance testing. Steve Koch (DTC, GSD) and Bob Gall (DTC, NCAR) provided very helpful guidance in coordinating DTC activities. Tom Schlatter and Steve Weygandt helped formulate the recommendations.

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA.

8. References

Benjamin, Stan, and John Brown, 2006: WRF Rapid Refresh Dynamic Core Evaluation and Recommendation. Memo to S. J. Lord and N. Seaman, 31 August 2006, 62pp. Available at http://ruc.noaa.gov/coretest2/WRF_Rapid_Refresh_dynamic_core_recommendation-31aug06.pdf

______, 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, 2004a: An hourly assimilation/forecast cycle: The RUC. *Mon. Wea. Rev.*, **132**, 495-518

______, G. A. Grell, J. M. Brown, and T. Smirnova, 2004b: Mesoscale weather prediction with the RUC hybrid isentropic-terrain-following coordinate model. *Mon. Wea. Rev*., **132**, 473-494.

______, D. Devenyi, T. Smirnova, S. Weygandt, J.M. Brown, S. Peckham, K. Brundage, T.L. Smith, G. Grell, and T. Schlatter, 2006: From the 13-km RUC to the Rapid Refresh. *12th Conf. on Aviation, Range, and Aerospace Meteorology (ARAM)*, Atlanta, Amer. Meteor. Soc.

______, S. Weygandt, J. M. Brown, T. Smirnova, D. Devenyi, K. Brundage, G. Grell, S. Peckham, T. L. Smith, T. Schlatter, and G. Manikin, 2007: From the radar-enhanced RUC to the WRF-based Rapid Refresh. *22nd Conf. Wea. Anal. and Fcstg./18th Conf. Numer. Wea. Pred.*, Park City, Amer. Meteor. Soc., Paper J3.4.

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.

Chen, F., K. Mitchell, J. Schaake, Y. Xue, H.-L. Pan, V. Koren, Q. Y. Duan, M. Ek, A. Betts, 1996: Modeling of land-surface evaporation by four schemes and comparisons with FIFE observations. *J. Geophys. Res.*, **101**(D3), 7251-7268.

Grell, G. A., and D. Devenyi, 2002: A generalized approach to parameterizing convection. *Geophys. Research Letters* **29**, 38-1-38-4.

Janjic, Z. I., 1994: The step-mountain eta-coordinate model: further development of the convection, viscous sublayer and turbulence closure schemes. *Mon. Wea. Rev*., **122**, 927-945.

______, 2001: Nonsingular Implementation of the Mellor-Yamada Level 2.5 Scheme in the NCEP Meso Model. National Centers for Environmental

Prediction Office Note 437, 61 pp. Available at http://www.emc.ncep.noaa.gov/officenotes/FullTOC.html#2000

______, and J. P. Gerrity, Jr., 2001: An alternative approach to nonhydrostatic modeling. *Mon. Wea. Rev.*, 129, 1164-1178.

Kain, John S., S. J. Weiss, J. J. Levit, M. E. Baldwin and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 307-324.

Koren, V., J. Schaake, K. Mitchell, Q.-Y. Duan, F. Chen, and J. M. Baker, 1999: A parameterization of snowpack and frozen ground intended for NCEP weather and climate models. *J. Geophys. Res.*, **104**(D16), 19569-19586.

Nance, Louisa Bogar, 2006: Weather Research and Forecasting Core Test: DTC Report. Developmental Testbed Center informal report, 36pp. Available at http://ruc.noaa.gov/coretest2/DTC_report.pdf

______, L. R. Bernardet, B. Weatherhead, G. Noonan, T. Fowler, T. G. Smirnova, S. G. Benjamin, J. Brown and A. Loughe, 2007: Weather Research and Forecasting Core Tests. *22nd Conf. Wea. Anal. and Fcstg./18th Conf. on Numer. Wea. Pred.*, Park City, Amer. Meteorol. Soc., Paper 5B.4.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. C. Powers, 2005: A Description of the Advavnced Research WRF Version 2. NCAR Technical Note NCAR/TN-468+STR, 88pp.

Smirnova, T.G., J.M. Brown, and D. Kim, 2000: Parameterization of cold-season processes in the MAPS land-surface scheme. *J. Geophys. Res.* **105**, 4077-4086.

Thompson, G, R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. Mon. Wea. Rev, 132, 519-542.

Weatherhead, E., and G. Noonan, 2006a: Statistical Analysis of Differences in Wind between ARW and NMM WRF Cores. Developmental Testbed Center informal report, 98pp.

______, and ______, 2006b: Statistical Analysis of Differences in Aircraft Temperature Data between ARW and NMM WRF Cores. Developmental Testbed Center informal report, 104pp.

______, ______, T. Fowler, L. Bernardet, L. Nance, and S. Koch, 2006a: Summaries of statistical analyses of differences in Relative Humidity, Temperature, Wind and Precipitation between NMM and ARW WRF Cores using both Sonde and Aircraft Data. Developmental Testbed Center informal report, 7pp. Available at http://ruc.noaa.gov/coretest2/Weatherhead-summary.pdf

______, ______, ______, ______, ______, and ______, 2006b: Statistical analyses of differences in Relative Humidity, Temperature, Wind and Precipitation between NMM and ARW WRF Cores. Developmental Testbed Center informal report, 276pp. Available at http://www-frd.fsl.noaa.gov/mab/wrfrr/doc/Report%201-statistical%20analyses.pdf

Weygandt, S. S. and S. G. Benjamin, 2007, Radar-reflectivity based initialization of precipitation systems using a diabatic digital filter initialization within the Rapid Update Cycle. *22nd Conf. Wea. Anal. and Fcstg./18th Conf. Numer. Wea. Pred.,* Park City, Amer. Meteorol. Soc., Paper 1B.7.