

P4.475 A COMPARISON OF WRF 3.2 AND WRF 3.1.1 FORECASTS OF EIGHT NOR'EASTER EVENTS

Stephen D. Nicholls* and Steven G. Decker

Department of Environmental Sciences, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

1. INTRODUCTION

Since 1950 when a team of Princeton scientists completed the first successful numerical weather forecast on ENIAC, numerical weather prediction has become an integral aspect of atmospheric science. As computer technology has advanced, so have the complexity and capability of prognostic atmospheric models. Present day models, such as the Weather Research and Forecasting Model (WRF) Advanced Research WRF (WRF-ARW; hereafter shortened to WRF), solve computationally intensive, fully compressible, non-hydrostatic equations at specified time steps and using specified boundary conditions, which allow simulations of the atmosphere at model resolutions of which the Princeton scientists could have only dreamed. Despite such advancements, the WRF model remains imperfect; thus, it is updated and refined annually as new errors are found and new schemes are implemented.

In its version 3.2 release, the WRF model underwent several modifications, including additional microphysics schemes, updates to the NOAH LSM, and a key update to the Rapid Radiation Transfer Model (RRTM) longwave radiation scheme (UCAR 2010). This RRTM update was needed because of an error in how WRF manages the buffer layer between the WRF model top and the top of the atmosphere. In RRTM 3.1, a single buffer layer was added between the model top and the top of the atmosphere. Within this layer, the temperature was assumed to be isothermal and the mixing ratio constant (if data is present, else it decreases with decreasing pressure). Using such assumptions, Cavallo et al. (2011) found a Global Forecasting System (GFS) longwave heating rate bias of -15K per day above 100 hPa and a slight warming near the tropopause, which limits model stability (See Figure 1). Similar biases were also demonstrated for the WRF model (not shown). In their research, Cavallo et al. (2011) proposed a solution to the RRTM cold bias by making three changes. These changes include adding additional layers to their buffer layer ($dp=4$ hPa in their work), assuming temperature in this layer follows the standard mean temperature profile, and finally assuming water vapor concentration in this layer to be fixed at 5 ppmv. With these changes, this

cold bias was successfully reduced to within ± 0.5 K day^{-1} of the standard cooling rates. Here, we will compare the performance of WRF version 3.1.1 to WRF version 3.2 during eight nor'easter simulations and also determine to what degree the new RRTM 3.2 longwave scheme has contributed to these performance differences.

2. METHODOLOGY

This WRF-based study will solely focus on intense wintertime cyclones that affect the Northeastern United States (nor'easters) given the complex interaction between dynamics and thermodynamics that underpin their development and maintenance. Specifically for this study, we chose eight nor'easter cases that occurred between the months of October – April and the years 2006 – 2010 (See Table 1). We chose this period because it was both consistent with the nor'easter season defined by Jacobs et al. (2005) and because these events correspond with a satellite dataset that we used in a parallel study. By using the entire season, we aim to diversify the number of unique weather situations where a nor'easter later develops.

2.1 WRF MODEL SETUP AND CONFIGURATION

Each of our WRF model runs was configured to allow it to simulate each cyclone from its initial genesis through its passage east of Newfoundland, Canada. To provide focus on the Northeastern United States (NEUS) and to make the best use of our computational resources, we chose a WRF model setup comprised of three nested domains (graphically shown in Figure 2) with two-way interaction and 27 vertical levels. Horizontal resolution on these three model domains were 45, 15, and 5 km for domains 1, 2, and 3, respectively. Our selected horizontal and vertical resolutions were selected to be consistent with our input and boundary condition files from the GFS 003 model forecast, which has 28 vertical levels and 1-degree horizontal resolution. Finally, we selected a model top of 50 hPa to avoid data assimilation errors at higher levels (e.g., Cucurull et al. 2008).

To meet our research objectives of comparing WRF model performance and the impact of the new RRTM 3.2 scheme, we ran 24 total model runs. Of these runs, 8 were control runs (WRF 3.1.1 with RRTM 3.1; hereafter just WRF 3.1.1), and 16 experimental runs

*Corresponding author address: Stephen D. Nicholls, Rutgers University, 14 College Farm Rd., New Brunswick, NJ, 08901; email: nicholls@envsci.rutgers.edu

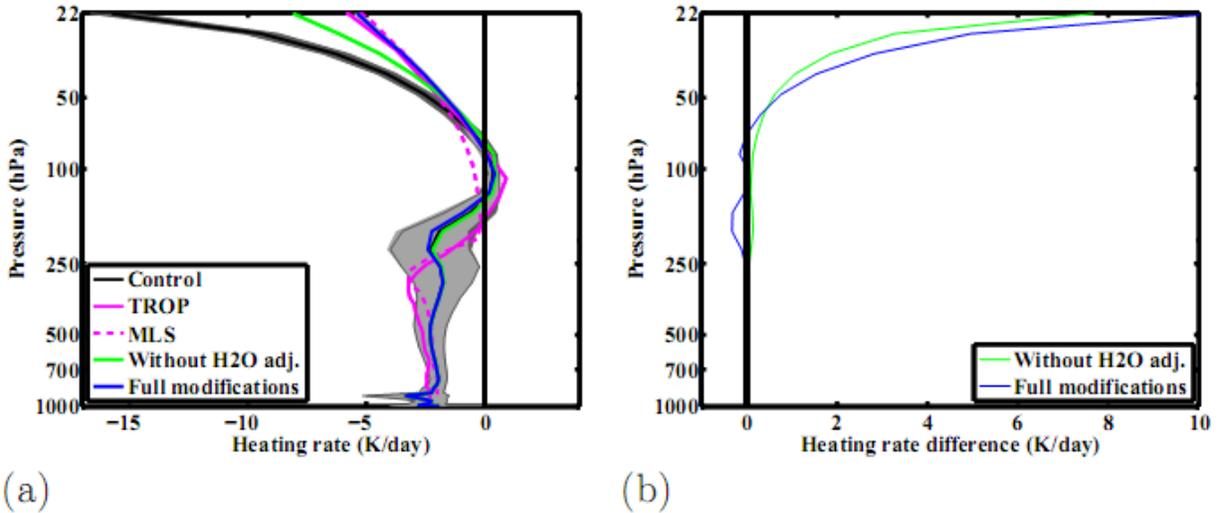


Figure 1: Composite vertical longwave potential temperature (a) heating rates and (b) differences from RRTM 3.1 control [experimental – control] for 6-hour GFS forecasts. ‘Control’ represents values from the unmodified RRTM 3.1 longwave radiation scheme, which assumes temperature to be isothermal above the model top and sharply decreases water vapor content with height. ‘Without H2O adj.’ modifies the original RRTM 3.1 scheme by fitting a mean vertical temperature profile above the model top. ‘Full modifications’ applies these same temperature modifications above the model top, but in addition fixes the water vapor content at 5 ppmv in this same region. ‘TROP’ and ‘MLS’ represent the standard tropical and mid-latitude summer standard atmospheric profiles, respectively. This figure is adapted from Cavallo et al. (2011), their Fig. 5.

Table 1: NESIS scores, date of impact in the Mid-Atlantic and Northeast United States, and model runtime specifics for the eight nor’easter cases studied. NESIS denotes the Northeast Snowfall Impact Scale (Kocin and Uccellini 2004).

Case Number	NESIS	Dates of Event	Model Start	Model End
1	N/A	22-24 Nov 2006	11/19 12Z	11/27 00Z
2	2.55	15-17 Mar 2007	3/12 18Z	3/20 06Z
3	N/A	15-17 Apr 2007	4/12 06Z	4/19 18Z
4	1.65	1-2 Mar 2009	2/26 12Z	3/6 00Z
5	N/A	15-16 Oct 2009	10/12 12Z	10/20 00Z
6	4.03	19-20 Dec 2009	12/16 06Z	12/23 18Z
7	4.3	4-7 Feb 2010	2/2 18Z	2/10 6Z
8	N/A	12-14 Mar 2010	3/9 06Z	3/16 18Z

involving 8 WRF 3.1.1 with RRTM 3.2 runs and 8 WRF 3.2 runs. Each of these model runs utilized identical WRF physics parameterization options. The only difference between them was the usage of the updated RRTM scheme in the “WRF 3.1.1 with RRTM 3.2” model run, and the WRF 3.2 run used all available model updates including RRTM 3.2.

Before running WRF, we also had to choose when to initialize our model runs and over what time duration. Given the dependence of nor’easter development on both upper-level vorticity advection (Jacobs et al. 2005) and pre-existing surface fluxes (Kuo et al. 1991, Ren et

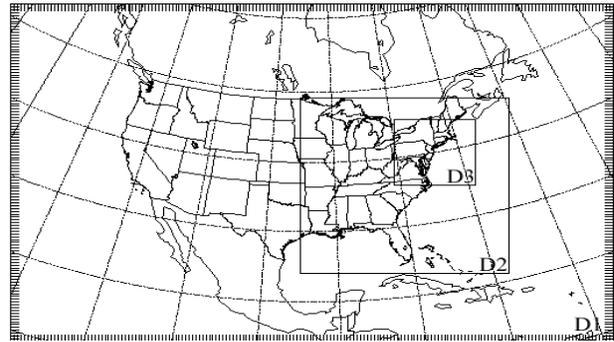


Figure 2: WRF-ARW model domains used in this study. The model resolutions for domain 1 (D1), domain 2 (D2), and domain 3 (D3) are 45, 15, and 5 km, respectively.

al. 2004), we initialized our runs prior to rapid cyclogenesis. Given our interest in the NEUS region, and specifically New Jersey, we chose to initialize our cases 72 hours prior to first 0.02” precipitation measurement recorded at any New Jersey station that could be attributed to a case. The 72-hour lead time was chosen for several reasons. First, Kuo et al. (1991) noted that sensible and latent heat fluxes were best able to affect the resulting cyclogenesis during its formative stages (about 24-48 hours prior to the onset of rapid cyclogenesis). Second, a more recent study by Mote et

al. (1997) successfully applied a 72-hour lead time in their heavy snowstorm composite analysis study. Finally, we argue that a 72-hour lead time provides WRF with enough ‘spin-up’ time, yet leaves sufficient forecast uncertainty to make simulated differences between our WRF runs more apparent. Given our lead time, we needed a 180-hour WRF simulation duration to cover each case fully. The specific model run times are noted in Table 1.

2.2 EVALUATION

Evaluation of WRF model performance in this research involved both an inter-WRF comparison and another to the GFS 003 model analysis, which we use a ground truth. To generate our ground truth dataset, GFS 003 model analysis data was pre-processed by WRF and then interpolated to our WRF model grid. Our decision to use the GFS 003 model analysis as our ground truth was based upon our usage of it to initialize our WRF model runs.

Next, we evaluated WRF model performance using two methods: storm track and the energy norm. Storm tracks for both our ground truth and WRF model runs were derived via a custom algorithm that utilizes sea-level pressure (SLP) fields. This algorithm determines storm tracks in three steps. First, it determines whether each grid point is the lowest sea-level pressure minimum within a five-degree radius of its location at a given time. If it is the minimum, its location and value is noted, else it is ignored. Second, the uniqueness of a candidate point is determined by comparing its position and time information to that of the last stored point in existing storm track arrays. Only if a candidate point was within five degrees distance and six hours to a previous storm position was it then appended to an existing storm track. Otherwise, it designated a new storm track. Finally, if a stored storm track was comprised of two or more points, we compared each new candidate point to an estimated storm position at the current time step rather than the last storm track position. We derived this position by simply adding the difference in distance between the last two points in a storm track array to the most recent storm position in that array. Given the nature of these data, we will only evaluate and validate the storm track qualitatively with respect to the location and timing of each track

For our main quantitative analysis, we elected to use the energy norm described in Rabier et al. (1996), which is shown below.

$$\langle \Delta x, \Delta x \rangle = -\frac{1}{2} \int_{p_{sf}}^{p_{top}} \int_A \left(\Delta u^2 + \Delta v^2 + R_d T_r * (\Delta \ln(p_s))^2 + \frac{c_p}{T_r} \Delta T^2 \right) dx dy dp \quad (1)$$

In equation 1, $\langle \Delta x, \Delta x \rangle$ is the energy norm ($J m s^{-2}$), u is the zonal wind ($m s^{-1}$), v is the meridional wind ($m s^{-1}$), R_d is the dry air gas constant, T_r is the mean surface temperature (K), P_s is the surface pressure (Pa), c_p is heat capacity at constant pressure, and T is air temperature (K). Our decision to apply the energy norm rather than the root mean square error (RMSE) for our quantitative comparison stems from the fact that RMSE merely evaluates a slice of the atmosphere, whereas the energy norm evaluates the entire model volume. Additionally, RMSE is typically applied for one variable at a time (e.g., 500-hPa geopotential height, 850-hPa temperature, etc), while in comparison, the energy norm calculates a combined error from four meteorological variables. Despite their mathematical differences, both RMSE and the energy norm can be interpreted similarly. For both indices, higher values denote greater error and values near zero denote an almost perfect forecast.

3. RESULTS

3.1 STORM TRACK ANALYSIS

The WRF simulated storm tracks and that of the ‘ground truth’ for all eight cases are shown in Figure 3. Looking solely at the ground truth (black lines), we were able to classify each event as a Type A or Type B cyclone using the Miller Classification System (Miller 1946). Under the Miller Classification System, a Type A cyclone typically develops over or near the Gulf of Mexico and then propagates up the eastern US seaboard. In contrast, a Type B cyclone propagates west of the Appalachian Mountains and then coastal redevelopment occurs later. Using these classifications, we determined that all of our cases sans the February 2010 case were Type A cyclones.

After assigning a Miller classification to each storm track, and recording both the minimum sea-level pressure and storm positions of the ground truth, we then proceeded to compare it to our WRF model simulations. In Figure 3, storm tracks for the WRF 3.1.1, WRF 3.1.1 with RRTM 3.2 and WRF 3.2 are denoted with the magenta, brown, and yellow lines, respectively. Starting with the Miller classification, it is clearly shown that all cases except March 2007 are of the correct type. In this 2007 case, all WRF runs clearly show a Type B cyclone track, whereas the ground truth is a Type A. This difference can be largely attributed to the failure of all three WRF simulations to correctly generate the surface low near the Gulf of Mexico and instead place it near the Great Lakes. One of the likely culprits is a sharper and deeper 500-hPa geopotential shortwave trough over Louisiana and Mississippi on March 15 in the ground truth that is not resolved by WRF (Figure 4, top panel). The sharpness of this trough is crucial

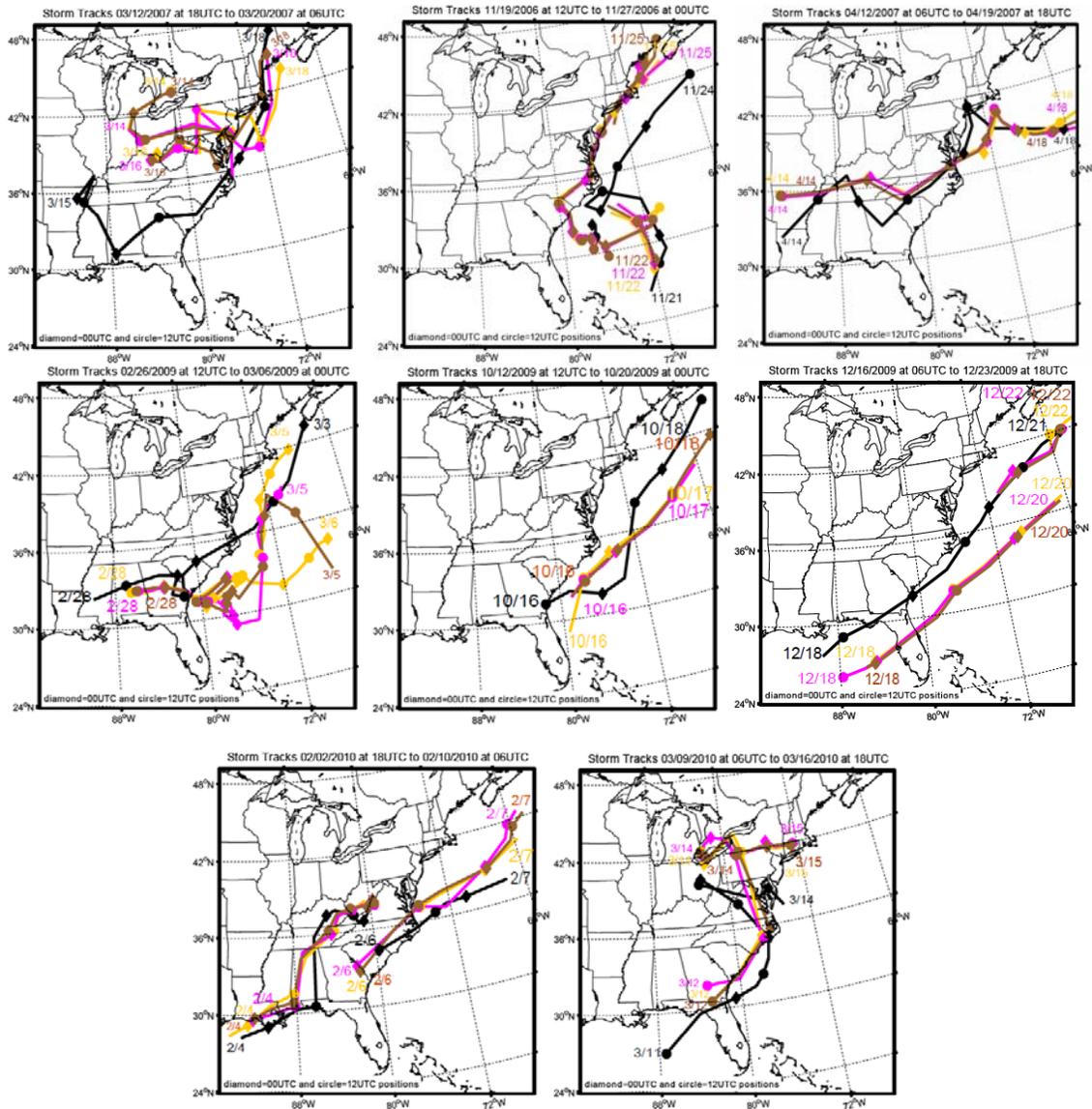


Figure 3: Sea-level pressure-based storm tracks for all eight nor'easter cases with the start date and end date of each track noted. Here we show the GFS 003 model analysis ground truth (black), WRF 3.1.1 unmodified (magenta), WRF 3.1.1 with RRTM 3.2 (brown), and WRF 3.2 (yellow).

because it enhanced upward vertical motion over the region where the actual surface low later developed.

Next, we looked at SLP. All WRF cases varied little (< 5 hPa) relative to each other and showed no clear and consistent intensification bias relative to the 'ground truth'. Specifically, we found WRF to under-intensify the main cyclone in 2 of 8 cases (October 2009 and December 2009). For these two cases WRF under-intensified the main cyclone by 5 and 20 hPa, respectively. In contrast, WRF over-intensified the

main cyclone in 3 of 8 cases (November 2006, April 2007, and February 2010) by 8, 8 and 6 hPa, respectively. All other cases were within 5 hPa and were deemed to be the correct intensity. Inter-WRF comparisons revealed that WRF 3.2 better calculated the minimum SLP in 5 of 8 cases relative to WRF 3.1.1 and 4 of 8 cases relative to WRF 3.1.1 with RRTM 3.2. Between the two WRF 3.1.1 runs, WRF 3.1.1 with RRTM 3.2 better simulated SLP in 5 of 8

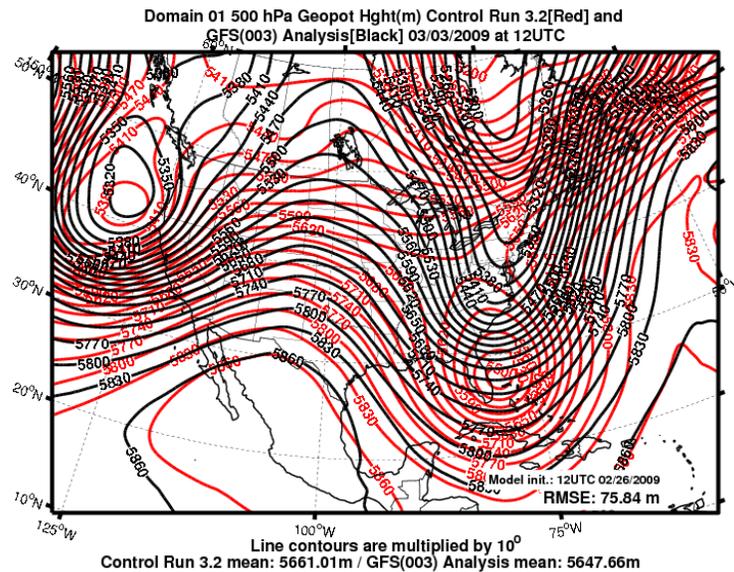
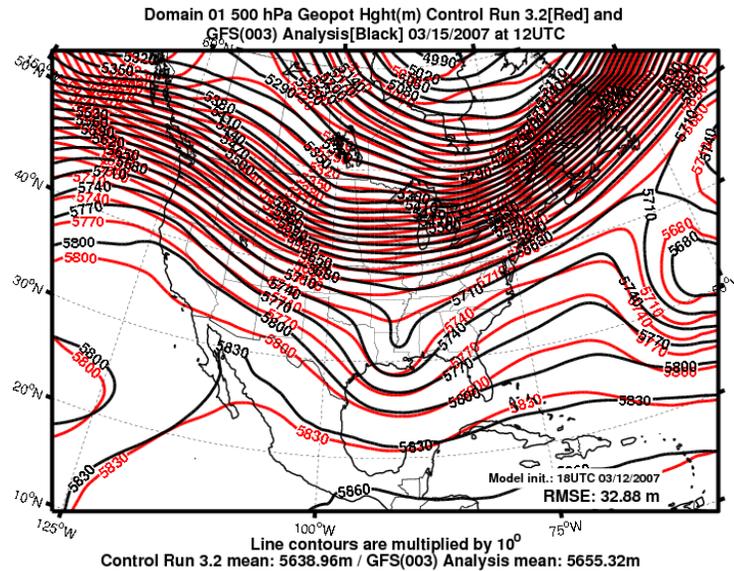


Figure 4: 500-hPa geopotential height from the ground truth (black lines) and the WRF 3.2 (Red lines) on 13 March 2007 (top) and 3 March 2009 (bottom).

cases. However, as mentioned above, these WRF-model SLP differences were relatively small.

Inter-comparing both the timing and positioning associated with WRF model storm tracks shows only small storm track discrepancies in all but two cases (March 2007 and February 2009). Storm tracks for the six low discrepancy cases strayed no more than 50 km from the other, whereas it was sometimes over 100 km in the other two cases. Despite these variations, all eight cases showed that WRF 3.2

consistently lead its competitors by less than six hours and also WRF 3.1.1 with RRTM 3.2 followed more closely to WRF 3.1.1 than WRF 3.2. Hence we can hypothesize that RRTM 3.2 indeed has an impact on the simulated storm track, but other model upgrades (such as the NOAA LSM) likely explain most of the observed variance between WRF 3.2 and WRF 3.1.1.

Using WRF-based SLP and 500-hPa geopotential height fields, we were able to show that the simulated

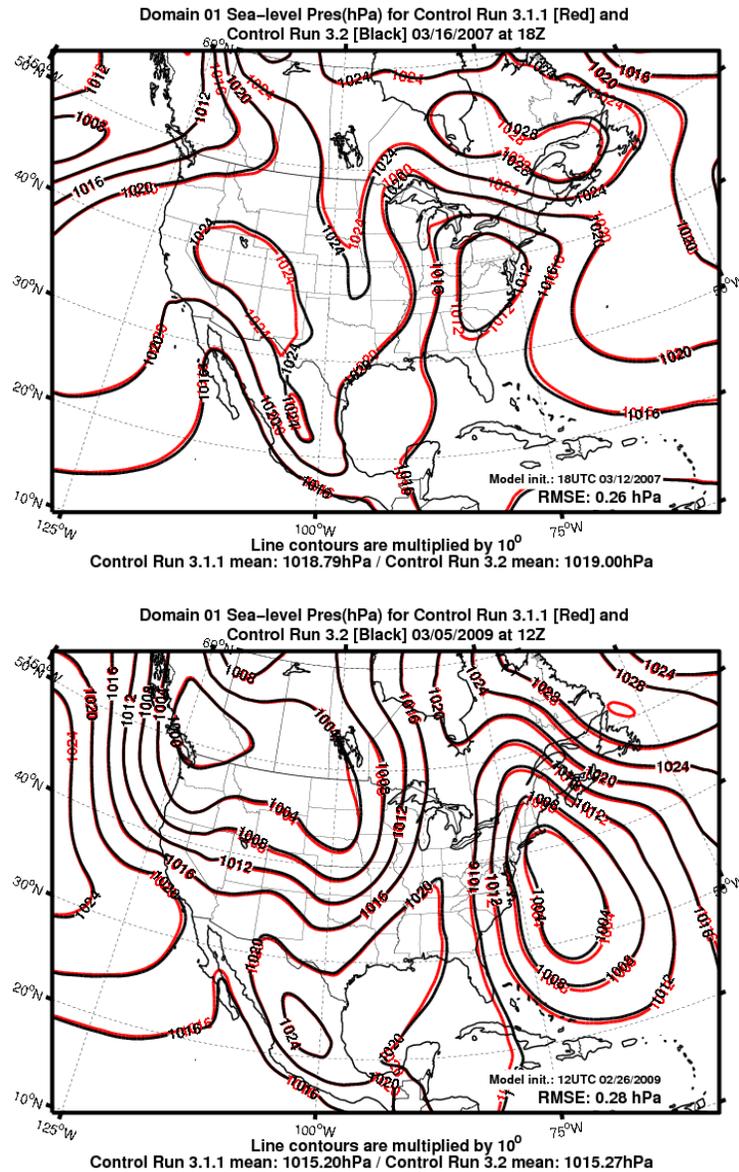


Figure 5: Sea-level pressure (hPa) from the WRF 3.2 (black lines) and the WRF 3.1.1 (Red lines) on 16 March 2007 (top) and 5 March 2009 (bottom).

large-scale environment varied little between WRF versions for any particular case (not shown for all cases). We can explain the higher track variance for the March 2007 and March 2009 cases from Figure 5. In this figure, we show SLP fields at 18 UTC 16 March 2007 and 12 UTC 5 March 2009 for WRF 3.2 and WRF 3.1.1. At these times, the storm track variance between these runs was over 50 km (see Figure 3),

but the RMSE values were small (0.26 and 0.28 hPa, respectively). Given these small RMSE values, the high track variance is most likely explained by SLP perturbations within the relative weak but expansive low-pressure regions rather than by notable WRF simulation differences from the various model updates.

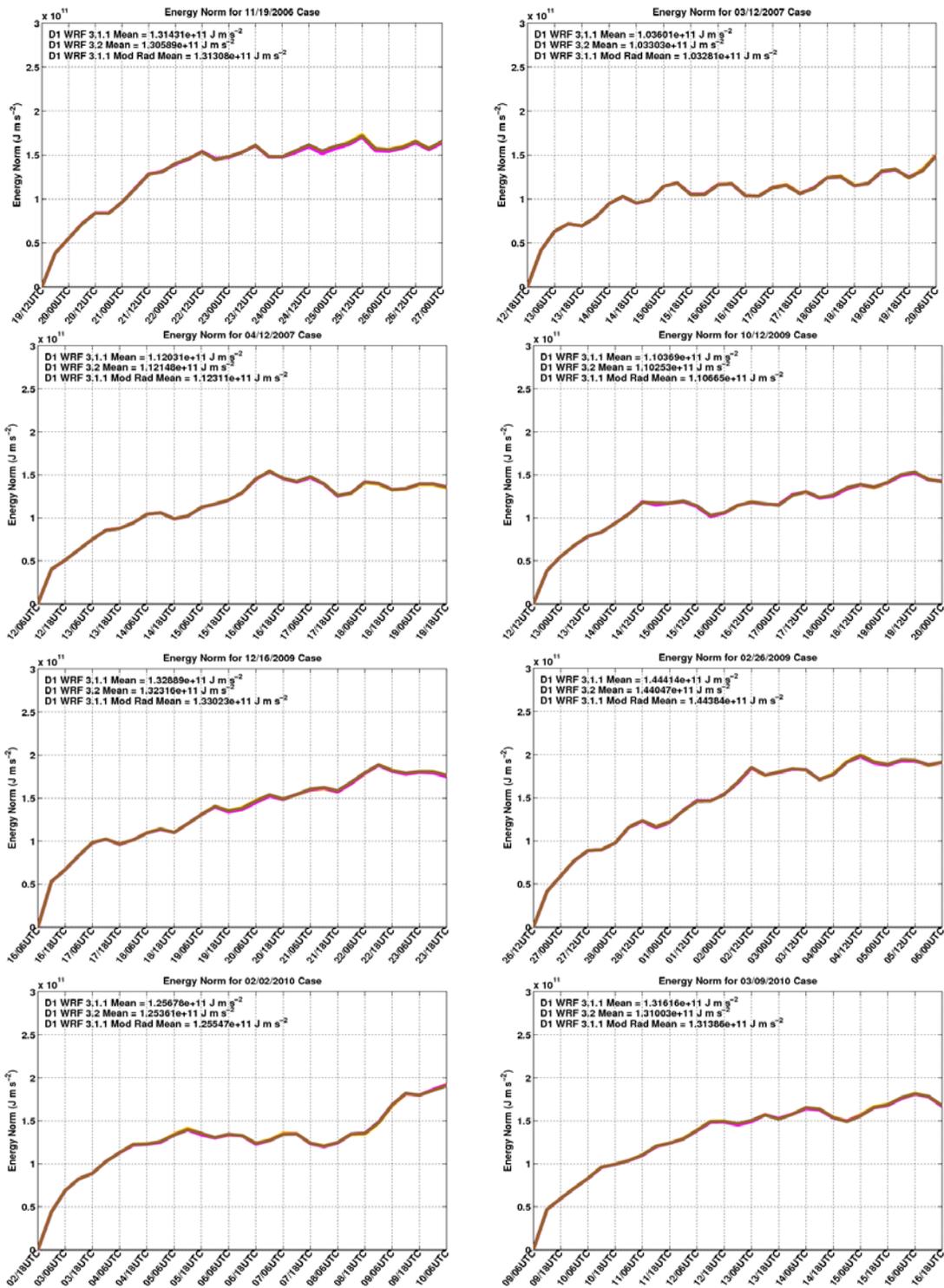


Figure 6: Energy norm for all eight nor'easter cases. Shown line colors denote WRF 3.1.1 (magenta), WRF 3.1.1 with RRTM 3.2 (brown) and WRF 3.2 (yellow)

Table 2: Energy norm p-values from two-tailed T-tests assuming unequal variance for each of our eight cases and overall.

Two-tailed T-test (P value)	11/19/2006	3/12/2007	4/12/2007	2/26/2009	10/12/2009	12/16/2009	2/2/2010	3/9/2010	All
WRF 3.2 / WRF 3.1.1	0.9381634	0.968568	0.989888	0.977987	0.9892487	0.9590012	0.97494	0.95568	0.92076
WRF 3.2 / WRF 3.1.1 RRTM 3.2	0.9471588	0.997708	0.986009	0.979794	0.9618558	0.9495918	0.9853	0.97234	0.92409
WRF 3.1.1 / WRF 3.1.1 RRTM 3.2	0.9910243	0.966278	0.975884	0.998204	0.9725796	0.9905099	0.98962	0.98334	0.99668

Table 3: Energy norm-based performance of WRF 3.1.1, WRF 3.2, and WRF 3.1.1 with RRTM 3.2 for our eight nor'easter cases and overall. Here we denote a win as one experimental model having the lowest energy norm value than its two counterparts. The upper table shows the number of time steps where each model wins and the lower panel show the corresponding percentage of time steps this represents.

Physical Count	11/19/2006	3/12/2007	4/12/2007	2/26/2009	10/12/2009	12/16/2009	2/2/2010	3/9/2010	All
WRF 3.1.1 Wins	2	5	10	7	12	7	7	1	51
WRF 3.2 Wins	22	14	19	19	16	22	15	26	153
WRF 3.1.1 RRTM 3.2 Wins	7	12	2	5	3	2	9	4	44
Total Obs	31	31	31	31	31	31	31	31	248

Percentage	11/19/2006	3/12/2007	4/12/2007	2/26/2009	10/12/2009	12/16/2009	2/2/2010	3/9/2010	All
WRF 3.1.1 Wins	6.45%	16.13%	32.26%	22.58%	38.71%	22.58%	22.58%	3.23%	20.56%
WRF 3.2 Wins	70.97%	45.16%	61.29%	61.29%	51.61%	70.97%	48.39%	83.87%	61.69%
WRF 3.1.1 RRTM 3.2 Wins	22.58%	38.71%	6.45%	16.13%	9.68%	6.45%	29.03%	12.90%	17.74%

Finally, we assessed how WRF-simulated storm tracks compared to the ground truth to determine any temporal or positional bias. Temporally, our WRF runs lagged behind the 'ground truth' in 4 of 8 cases (November 2006, February 2009, December 2009 and February 2010) and lagged by 24, 48, 24, and 12 hours, respectively. In contrast, only for April 2007 did WRF lead the 'ground truth' by 12 hours. The remaining cases were within 6 hours of their ground truth counterparts and were considered on time. In addition to these time biases, WRF more often than not (5 of 8 cases) had a leftward storm track bias relative to the ground truth. We hypothesize that these spatial and temporal differences are due to three causes. First, all our WRF model runs were cold-start runs, and therefore required 6-12 hours to spin up. Such a spin-up may cause a systematic time lag because existing weather features may develop more slowly in WRF than in the ground truth. Second, our WRF boundary conditions files were generated from GFS 003 model forecasts initialized at the times specified in Table 1. Therefore, our model simulations were influenced by GFS model errors propagating through our domains. Third, our WRF simulations may mishandle the development of synoptic and

mesoscale weather features, which are crucial to the development of each case.

This last situation is best illustrated in Figure 4, where 500-hPa heights are shown for WRF 3.2 (red lines) and the ground truth on March 15, 2007 at 12 UTC (March 2007 case) and March 3, 2009 at 12 UTC (February 2009 case). We selected these two cases and times because they illustrated how different the WRF simulations and ground truth became. As seen in Figure 4, the ground truth on March 15 shows a sharp, mesoscale height trough that is not represented by WRF. This trough is crucial because the actual surface low developed just eastward of it a short time later and therefore its existence and associated upward vertical motion can be at least partially attributed to its presence. For the March 3 case, WRF 3.2 develops a deep cut-off height minimum, which does not happen in the ground truth. After this time, the ground truth trough quickly moves up the eastern seaboard, whereas the cut-off height minimum meanders near Florida for two days. As a result, all WRF model storm tracks eventually lagged the ground truth by 48 hours.

Table 4: Energy norm-based performance for WRF 3.2 versus WRF 3.1.1 and WRF 3.1.1 with RRTM 3.2 versus WRF 3.1.1. A 'win' is counted if an experimental run has a lower energy norm than WRF 3.1.1.

WRF 3.1.1 RRTM 3.2 vs WRF 3.1.1	11/19/2006	3/12/2007	4/12/2007	2/26/2009	10/12/2009	12/16/2009	2/2/2010	3/9/2010	All
Physical Count (RRTM 3.2 Wins)	17	19	9	18	8	20	17	22	130
Percentage	54.84%	61.29%	29.03%	58.06%	25.81%	64.52%	54.84%	70.97%	52.42%
WRF 3.2 vs WRF 3.1.1	11/19/2006	3/12/2007	4/12/2007	2/26/2009	10/12/2009	12/16/2009	2/2/2010	3/9/2010	All
Physical Count (WRF 3.2 Wins)	25	25	20	21	17	23	23	30	184
Percentage	80.65%	80.65%	64.52%	67.74%	54.84%	74.19%	74.19%	96.77%	74.19%

3.2 ENERGY NORM ANALYSIS

Figure 6 shows our energy norm results between each WRF model run and the ground truth. In this figure, the energy norms for WRF 3.1.1, WRF 3.1.1 with RRTM 3.2, and WRF 3.2 are shown with magenta, brown and yellow lines, respectively. We also provide the mean energy norm value for each WRF simulation at the top of each case panel. Upon quick observation of Figure 6, we noted three key findings. First, all WRF energy norms showed a clear diurnal signal and generally increased with time. Second, all WRF model runs are nearly indistinguishable from each other. Finally, WRF 3.2 generally performed better (had a lower energy norm) than both WRF 3.1.1 runs and WRF 3.1.1 with RRTM 3.2 performed marginally better than WRF 3.1.1.

The increasing energy norm and its diurnal signal were striking features on Figure 6. Energy norm values typically increase with time due to accumulating model errors resulting from increasing differences relative to the ground truth as model run time increases. What came as more of a surprise to us was the very apparent diurnal signal in each WRF simulation. Originally, we hypothesized that this signal may have been due in part to how WRF handled radiation, which served as a partial motivation for the creation of our WRF 3.1.1 with RRTM 3.2 runs. As shown in Figure 6, these model runs had the same diurnal cycle and therefore RRTM was not the cause. Ultimately, we broke down the energy norm into its components and found the diurnal cycle originated predominately from the temperature term (not shown). It should also be noted that the surface pressure term did contribute to this diurnal signal, but its contribution to the total energy norm was an order of magnitude less than temperature.

The raw energy norms shown in Figure 6 are nearly identical and their differences are not statistically significant. To prove this non-significance, we calculated p-values on our energy norm data

using a two-tailed Student T-test assuming unequal variance and a 0.05 significance level. As shown in Table 2, no one WRF model run achieved statistical significance relative to its counterparts. This is best proven by the 0.938 p-value between WRF 3.2 and WRF 3.1.1 in the November 2006 case. This is the smallest p-value on Table 2. The lack of statistical significance is not disastrous to this study considering that present day prognostic models have been shown to be capable of forecasting for up to several days with considerable skill. Hence had our results shown statistical significance, it would likely signal a critical error rather than a revolutionary improvement in our WRF simulations. Table 3 summarizes the energy norm-based performance of each model run. Overall, WRF 3.2 had the lowest energy norm in 153 (61.7%) of 248 total time steps and was the best simulation outright (>50% 'win' rate) in 7 of 8 cases. In individual model comparisons (Table 4), WRF 3.2 bested WRF 3.1.1 in 184 (74.19%) of 248 time steps and performed better overall in all 8 cases. Comparatively, WRF 3.1.1 with RRTM 3.2 had a lower energy norm in 130 (52.4%) of 248 cases, and performed better overall relative to WRF 3.1.1 in 6 of 8 cases.

To further visualize how WRF model performance varied with time, we generated plots of the difference in energy norm for each case in Figure 7. The shown differences are WRF 3.1.1-WRF 3.2 (magenta) and WRF 3.1.1-WRF 3.1.1 with RRTM 3.2 (brown). This differencing order was chosen because it makes positive values indicate improvement over WRF 3.1.1 which we argue is more intuitive. Each case has only subtle differences in the first 48 hours and then afterward these differences vary widely from case to case. The sudden spike in the energy norm differences after 48 hours stems from the subtle differences in various meteorological fields associated with each nor'easter case and other weather systems present in our model domain. Overall, WRF 3.2 generally improved our nor'easter simulations averaging a $3.7571 \times 10^8 \text{ J m s}^{-2}$ improvement in the

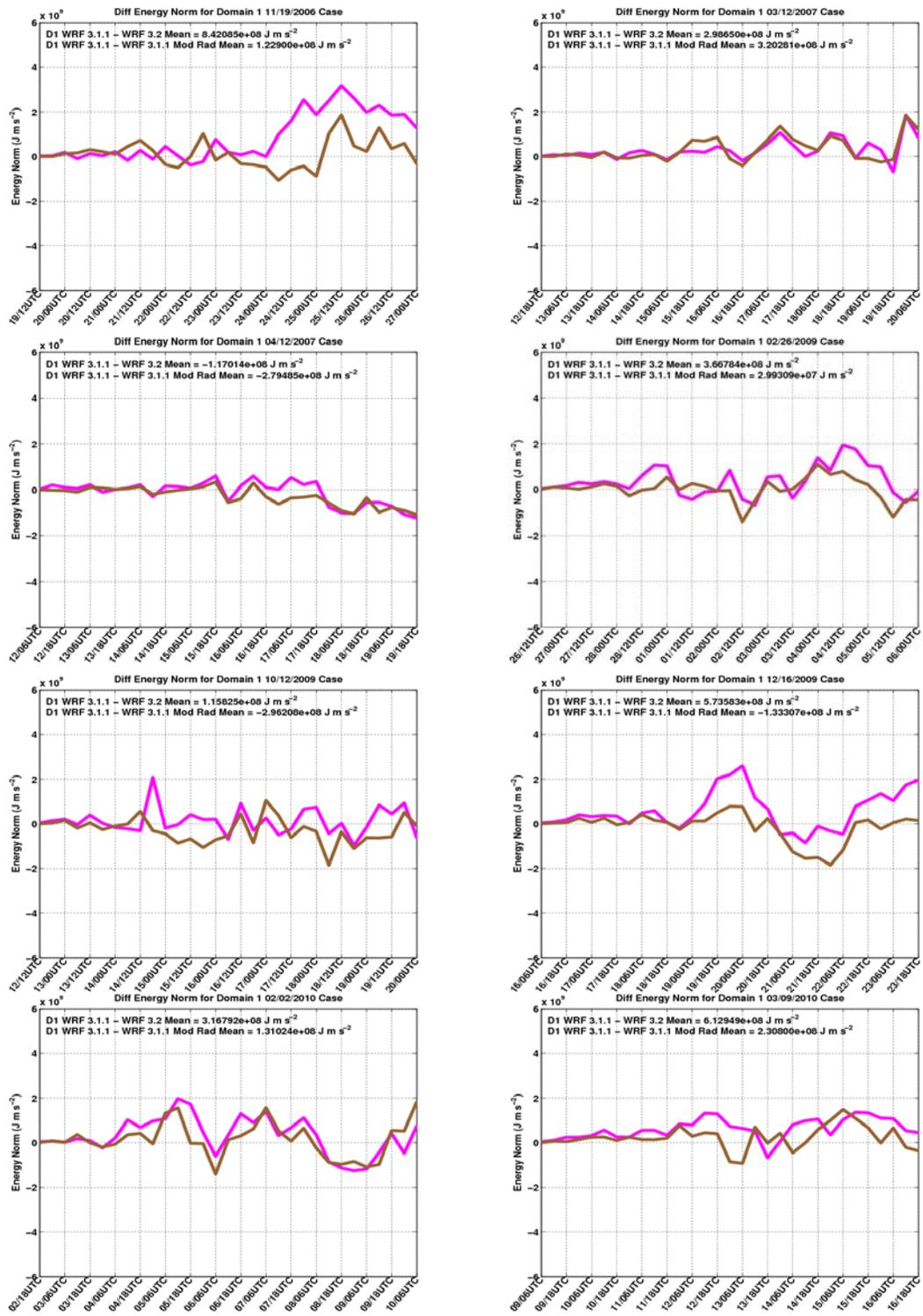


Figure 7: Difference in energy norm for all eight nor'easter cases. Shown line colors denote WRF 3.1.1 - WRF 3.2 (magenta) and WRF 3.1.1 - WRF 3.1.1 with RRTM 3.2 (brown)

energy norm. In comparison, WRF 3.1.1 with RRTM 3.2 averaged a smaller $1.56 \times 10^7 \text{ J m s}^{-2}$ improvement, which reflects its mixed performance results. During the main nor'easter event period (72-120 hours) WRF 3.2 is rarely negative and more often than not more positive than WRF 3.1.1 with RRTM 3.2. In contrast, WRF 3.1.1 with RRTM 3.2 tends to be either slightly positive or negative. With these results, we have shown that WRF 3.2 is a clear improvement over its predecessor, and this improvement cannot be solely attributed to implementation of RRTM 3.2.

For sake of completeness and comparison to our energy norm results, we also completed a RMSE-based analysis between WRF 3.2 and WRF 3.1.1 (see Table 5). In Table 5, we indicate the number of instances where the WRF 3.2 RMSE values of SLP, 850-hPa temperatures, and 500-hPa geopotential heights for each case, domain, and overall were lower than those from WRF 3.1.1. Most striking to us was the high variability of the WRF 3.2 performance between the different cases and variables. As an example, for the March 2007 case, WRF 3.2 has a lower 850-hPa temperature RMSE than WRF 3.1.1 on domain 1 in 31 (100%) of 31 time steps, whereas for the December 2009 case, it was only 8 (25.8%) of 31 time steps. These results differ from our energy norm analysis, which were generally more in favor of WRF 3.2. Specifically, WRF 3.2 had a lower energy norm than WRF 3.1.1 in 25 (80.65%) of 31 time steps and 23 (74.19%) of 31 time steps for these same two cases, respectively. The volatility of our RMSE analysis likely originates from the relatively small differences (less than 0.1 K on average) between the WRF model runs. Because of these small differences, even a minute change in the derived temperatures could easily change the outcome. Similar small differences were also noted for the 500-hPa geopotential height and SLP analyses where the average difference was less than 2m and 0.2 hPa, respectively. We hypothesize that the notable discrepancy between our RMSE and energy norm results are due to both the scale of these RMSE differences and also because RMSE is a 2D (x,y) quantity, whereas the energy norm is 3D (x,y,z). Overall, our RMSE analysis on domain 1 (most comparable to the energy norm) shows WRF 3.2 to provided superior simulations of SLP, 850-hPa temperatures, and 500-hPa geopotential heights for 140 (56.5%), 158 (63.7%), and 109 (49.8%) of 248 total time steps, respectively. Taking each variable separately WRF 3.2 forecasted better SLP fields in 4 of 8 cases, better 850-hPa temperatures in 6 of 8 cases, and better 500-hPa geopotential heights in 5 of 8 cases. Hence even though the RMSE results are

not as definitive, it still indicates that WRF 3.2 is in generally the superior model.

4. SUMMARY AND CONCLUSIONS

We assessed WRF model performance between versions 3.2 and 3.1.1 during eight nor'easter cases that occurred between October and April of 2006 to 2010. Between these versions, a notable upgrade was made to the RRTM longwave scheme to address a -15 K day^{-1} longwave heating bias that was present in RRTM 3.1 (Cavallo et al. 2011). To fully assess the performance differences between WRF 3.1.1 and WRF 3.2, as well as that from the addition of RRTM 3.2, we ran three WRF simulations for each case (WRF 3.2, WRF 3.1.1, and WRF 3.1.1 with RRTM 3.2). Each of these cases was identical, other than the addition of RRTM 3.2 to the WRF 3.1.1 with RRTM 3.2 model run, and all version 3.2 upgrades in the WRF 3.2 model run.

We evaluated the various WRF model runs to each other and to the ground truth via SLP-based storm track analysis and the calculated energy norm. Our inter-WRF comparison revealed that the various WRF run storm tracks rarely strayed any more than fifty kilometers from each other and their simulated cyclones were of similar intensity. Where these runs diverged further, the region of lowest pressure was expansive and relatively high. This allowed for subtle surface fluxes changes to alter the location of minimum sea-level pressure within this area. As compared to the 'ground truth', our WRF runs showed no clear and consistent cyclone intensification bias given that 2 of 8 cases were over-intensified and 3 of 8 cases were under-intensified. We did however note a time lag in 4 of 8 cases and a leftward storm track bias in 5 of 8 cases. These biases resulted from differences in synoptic and mesoscale dynamical fields, most notably geopotential height.

Our energy norm analysis failed to reveal any statistically significant differences amongst the various WRF model runs. This lack of statistical significance is not a disaster, but instead reflects the quality of the WRF model. We did find clear evidence that WRF 3.2 has generally superior performance to WRF 3.1.1 given its average $3.7571 \times 10^8 \text{ J m s}^{-2}$ reduction in the energy norm and overall better forecast in 7 of 8 cases. More specifically, WRF 3.2 had a lower energy norm and therefore lower forecast error in 184 of 248 time steps (74.19%). In contrast, the performance of WRF 3.1.1 with RRTM 3.2 is more mixed, but still generally positive. It averaged a $1.56 \times 10^7 \text{ J m s}^{-2}$ reduction in the energy norm, and had a lower energy norm in 130 of 248 time steps (52.4%). Hence, RRTM 3.2 just tends to slightly more

Case	500 hPa Geopotential Height (m)		850-hPa Temperature (K)		Sea-level Pressure (hPa)	
11/19/2006	Total obs per box 31					
<i>Domain 1</i>	20	64.52%	24	77.42%	20	64.52%
<i>Domain 2</i>	18	58.06%	24	77.42%	15	48.39%
<i>Domain 3</i>	23	74.19%	20	64.52%	17	54.84%
3/12/2007	Total obs per box 31					
<i>Domain 1</i>	29	93.55%	31	100.00%	27	87.10%
<i>Domain 2</i>	29	93.55%	26	83.87%	26	83.87%
<i>Domain 3</i>	23	74.19%	26	83.87%	26	83.87%
4/12/2007	Total obs per box 31					
<i>Domain 1</i>	19	61.29%	25	80.65%	3	9.68%
<i>Domain 2</i>	19	61.29%	27	87.10%	3	9.68%
<i>Domain 3</i>	23	74.19%	23	74.19%	18	58.06%
2/26/2009	Total obs per box 31					
<i>Domain 1</i>	24	77.42%	19	61.29%	16	51.61%
<i>Domain 2</i>	26	83.87%	18	58.06%	23	74.19%
<i>Domain 3</i>	22	70.97%	18	58.06%	14	45.16%
10/12/2009	Total obs per box 31					
<i>Domain 1</i>	23	74.19%	22	70.97%	21	67.74%
<i>Domain 2</i>	20	64.52%	22	70.97%	19	61.29%
<i>Domain 3</i>	21	67.74%	24	77.42%	15	48.39%
12/16/2009	Total obs per box 31					
<i>Domain 1</i>	6	19.35%	8	25.81%	11	35.48%
<i>Domain 2</i>	2	6.45%	11	35.48%	22	70.97%
<i>Domain 3</i>	7	22.58%	11	35.48%	16	51.61%
2/2/2010	Total obs per box 31					
<i>Domain 1</i>	6	19.35%	11	35.48%	4	12.90%
<i>Domain 2</i>	4	12.90%	14	45.16%	10	32.26%
<i>Domain 3</i>	12	38.71%	15	48.39%	13	41.94%
3/9/2010	Total obs per box 31					
<i>Domain 1</i>	13	41.94%	18	58.06%	7	22.58%
<i>Domain 2</i>	17	54.84%	23	74.19%	10	32.26%
<i>Domain 3</i>	23	74.19%	21	67.74%	15	48.39%
By Domain	Total obs per box 248					
<i>Domain 1</i>	140	56.45%	158	63.71%	109	43.95%
<i>Domain 2</i>	135	54.44%	165	66.53%	128	51.61%
<i>Domain 3</i>	154	62.10%	158	63.71%	134	54.03%
Overall	Total obs per box 744					
<i>All cases</i>	429	57.66%	481	64.65%	371	49.87%

Table 5: WRF 3.1.1 control versus WRF 3.2 RMSE results for sea-level pressure, 850-hPa temperature, and 500-hPa geopotential height for each case (31 time steps), domain (248 total time steps) and overall (744 total time steps). Above numbers and percentages reflect the instances where WRF 3.2 had a lower RMSE than WRF 3.1.1 control.

help than hurt the WRF 3.1.1 model simulations. Therefore, the improvement shown in WRF 3.2 can only be partially attributed to the introduction of RRTM 3.2. More likely the updates to the WRF-ARW dynamical core and other various model upgrades played a more significant role.

5. REFERENCES

Cavallo, S. M., J. Dudhia, and C. Snyder, 2011: A multi-layer upper boundary condition for longwave radiative flux to correct temperature biases in a mesoscale model. *Mon. Wea. Rev.*, in press.

Cucurull, L., J. C. Derber, R. Treadon, and R. J. Purser, 2008: Preliminary impact studies using Global Positioning System radio occultation profiles at NCEP. *Mon. Wea. Rev.*, **136**, 1865-1877.

Jacobs, N. A., G. M. Lackmann, and S. Raman, 2005: The combined effects of Gulf Stream-induced baroclinicity and upper-level vorticity on U.S. East Coast extratropical cyclogenesis. *Mon. Wea. Rev.*, **133**, 2494-2501.

Kocin, P. J., and L. W. Uccellini, 2004: A snowfall impact scale derived from Northeast storm snowfall distributions. *Bull. Amer. Meteor. Soc.*, **85**, 177-194.

Kuo, Y-H., R. J. Reed, and S. Low-Nam, 1991: Effects of surface energy fluxes during the early development and rapid intensification stages of seven explosive cyclones in the western Atlantic. *Mon. Wea. Rev.*, **119**, 457-476.

Miller, J. E., 1946: Cyclogenesis in the Atlantic coastal region of the United States. *J. Meteor.*, **3**, 31-44.

Mote, T. L., D. W. Gamble, S. J. Underwood, M. L. Bentley, 1997: Synoptic-scale features common to heavy snowstorms in the Southeast United States. *Wea. Forecasting*, **12**, 5-23.

Rabier, F., E. Klinker, P. Courtier, and A. Hollingsworth, 1996: Sensitivity of forecast errors to initial conditions. *Quart. J. Roy. Meteor. Soc.*, **122**, 121-150.

Ren, X., W. Perrie, Z. Long, and J. Gyakum, 2004: Atmosphere-ocean coupled dynamics of cyclones in the midlatitudes. *Mon. Wea. Rev.*, **132**, 2432-2451.

UCAR, 2010: WRF model version 3.2 updates. [Available online at <http://www.mmm.ucar.edu/wrf/users/wrfv3.2/updates-3.2.html>].

6. ACKNOWLEDGEMENTS

We express our gratitude to Dr. Steve Cavallo at the National Center for Atmospheric Research (NCAR) for his assistance and guidance with setting up our modified WRF 3.1.1 RRTM 3.2 model runs. Additionally, we thank NCAR for its management and distribution of the WRF-ARW model. Finally, I (SN) thank Rutgers University for its financial support of my teaching assistantship and its contribution towards attending the 2011 AMS Annual Meeting.