Turbulence PDT (TPDT) WRF ARW/NMM core comparisons

- Robert Sharman, Rod Frehlich, NCAR/RAL.
- Updated 8 September 2006

1. Executive summary

The TPDT examined the performance of turbulence metrics derived from both WRF candidate cores provided by the DTC. The performance was evaluated both qualitatively and quantitatively, for both 12 and 24 hour forecasts. We found that the WRF NMM core was definitely smoother than the ARW for all turbulence metrics computed. This did not translate however into better statistical performance as might be expected. In fact almost all turbulence diagnostics performed better using the WRF ARW core. We found that the performance differences were statistically significant and favor the WRF ARW over the WRF NMM. We thus strongly recommend adoption of the WRF ARW core in the operational RR-WRF model.

2. Comparison strategy

The WRF ARW/NMM core comparison strategy was based on the effect of various turbulence diagnostics that we have found from past experience to provide some skill for turbulence forecasting. These diagnostics were computed using the same code as is currently used in the FAA AWRP GTG2 (Graphical Turbulence Guidance) experimental product (now approved at the AWTT D3 level, expected to be approved at the D4 level in September). Most of the diagnostics are based on empirically-derived linkages between large scale processes and turbulence observations. Others are rooted in turbulence theory; one of them in particular uses 3d isotropic turbulence assumptions to predict eddy dissipation rates (EDR) computed from second-order structure functions of model velocity. We have found in the past that structure functions are very sensitive to resolved fine scale features in the model predicted velocity fields, so these were examined in some detail. The comparison strategy thus followed two lines of investigation:

- Use second-order structure functions to diagnose model smoothing and filtering effects
  - Less filtering is desirable for turbulence feature extraction
- Assess turbulence diagnostic statistical performance
  - Compute turbulence diagnostics (~30) from the current GTG suite of algorithms
  - Many of these diagnostics depend on first and second-order derivatives which accentuate differences in the underlying models
  - Compare contours of diagnostics (qualitative)
  - Compare to turbulence observations (PIREPs)
  - Develop POD curves => the higher the curve the better the overall performance (quantitative)
  - Assess the statistical significance of the differences in the areas under the curves

3. Results of structure function analyses

A quantitative measure of the effective spatial resolution of numerical models can be produced from the spatial spectrum or the spatial structure function (related to the spatial correlation function). A structure function of a velocity component x is defined as the average of the squared difference of x as a function of the separation distance or lag s. The structure function has the advantage of a robust theoretical interpretation at small lags (no contamination from aliasing which plagues interpretation of spectra) and permits useful local estimates of the small scale turbulence which are important for turbulence forecasts. (see e.g., Frehlich and Sharman 2004). A total of four structure functions can be derived from the longitudinal (velocity component is parallel to the separation vector) and transverse functions (velocity component is transverse to the separation vector) considering both east-west and north-south directions. Figs. 1-3 show some sample results from the analyses of the structure functions from the two candidate WRF cores. Some observations of the figures:

- The longitudinal and transverse structure functions of the horizontal velocity components u and v derived from the RUC13 (6 hr forecast), WRF ARW, and WRF NMM (6 and 12 hr forecasts) in Fig. 1 show lower levels at small separations in the WRF NMM than either the RUC13 or WRF ARW. This result implies that the horizontal velocities are slightly smoother in NMM than both the ARW and RUC13 models.
- The structure functions indicate that the ARW effective filtering is approximately the same as RUC13 effective filtering and less than the NMM effective filtering at altitudes of 5 &10 km. There is also no indication of uncorrelated model noise in any model which would appear as a constant level at small lags, i.e., the smallest resolvable scales in the model output contain information and are not “noise”.
- The transverse structure functions of vertical velocity w derived from RUC13 (6 hr forecast), WRF ARW and WRF NMM (6 and 12 hour forecast) in Fig. 2 show a) more small-scale energy in RUC13 compared to both WRFs, b) a consistent level at larger lags between the RUC13 and WRF ARW, and c) lower levels and pronounced filtering at all resolved scales from the WRF NMM compared to either the RUC13 or the WRF ARW. Overall, w is much smoother in WRF NMM than in WRF ARW, which would lead to underestimates of turbulence levels.
- There is also a slight enhancement of longitudinal structure functions in the EW direction which is most pronounced over Western US (Fig. 3). We believe this is produced by the enhanced gravity wave production and maintenance over the western mountain ranges.

4. Results of turbulence metric analyses

The results presented in this section are derived from 12h and 24h forecasts for all months in the DTC comparison exercise. Examples of the turbulence diagnostics produced by the models are shown in Fig. 4, which compares two of the better performing turbulence diagnostics derived from 12 hr forecasts of the RUC13, WRF ARW, and WRF NMM for a particular case. The thresholds used

for contouring the diagnostics are the same in each panel. For these thresholds the general pattern derived from all models is similar, but there are obvious differences in details.

- The RUC13 pattern is slightly larger than either WRF model.
- It is obvious by inspection that the WRF NMM is “smoother”, i.e. less detailed in the small scale features, for both diagnostics displayed. This enhanced smoothing in the WRF NMM is general and is apparent in plots of all 30 diagnostics. This difference is consistent with the statistical behavior seen in the structure function analyses.

Quantitative assessments of turbulence forecast performance were also performed using the techniques described in Sharman et al. (2006). The technique used was to compute the full suite of turbulence diagnostics within the FAA AWRP turbulence forecasting algorithm (GTG), together with a combination of 10 indices derived from each forecast model. The values of the diagnostics and the GTG combination were compared to available turbulence PIREPs within 1 ½ hours of the forecast time to derive probabilities of detection of moderate-or-greater (MOG) turbulence, referred to as PODY, and probabilities of detection of smooth or null turbulence, termed PODN. Obviously, both PODY and PODN approaching 1 are desirable. One method for assessing both PODY and PODN performance together is to construct curves of PODY vs PODN (or 1-PODN) by varying the MOG threshold values through ranges appropriate for each diagnostic. The area under these curves (AUC) can be used as a performance metric – larger AUCs indicating better discrimination of MOG events from null events. This process of diagnostic performance evaluation is detailed in Sharman et al. (2006).

Fig 5 compares pod curves of the 30 diagnostics, derived from RUC13, WRF ARW, and WRF NMM for 12 h forecasts for all months in the DTC comparison exercise. The 30 diagnostics are described in Sharman et al. (2006). These exhibit a spectrum of performance, from diagnostics showing very little skill (compare to the no-skill diagonal line, AUC=0.5) to higher skills with AUCs approaching 1. But quantitative comparisons are difficult to infer from these plots, so tables were constructed which display the AUCs for the different diagnostics derived from the 3 different models. Each table compares the top 20 diagnostics for each model using model 12h forecasts (Table 1) and 24 h forecast (Table 2) based on all 4 months computed in the DTC exercise. These tables also include the GTG combination of the top 10 indices for each model. The combination is based on a weighted sum of the 10 diagnostics with the weights proportional to the AUC^2, consistent with the current GTG formulation based on climatological performance (Sharman et al. 2006).

Some observations from inspection of the tables:

- Overall, the performance for diagnostics and the GTG combination as measured by the AUCs derived from the WRF ARW is better than from the WRF NMM for both 12 and 24 hour forecasts.

- The difference in performance is accentuated for most of the diagnostics in the 24 h forecasts.
- For the top 10 indices used in the GTG combination, the AUCs were ALL higher for the WRF ARW than for the WRF NMM, thus the GTG combination was also higher.
- The top 20 were not the same for both models but of the top 20, 18 were the same. Of these 18, 16 and 17 had higher AUCs using the WRF ARW than using the WRF NMM, for the 12h and 24 h forecasts respectively.
- The GTG combination AUCs were about 2%, 4% higher using the WRF ARW core than using WRF NMM core, for the 12, 24 h forecasts respectively.
- The edr diagnostic, which is based on second-order structure functions, has one of the largest differences between the WRF NMM and WRF ARW, the ARW based diagnostic metric (AUC) being about 5% higher for the 24 h forecast.

The statistical significance of these results is difficult to evaluate because of the limited number of observations (about 7,000 PIREPs) and the correlated nature of atmospheric processes. However, by random subsampling many times using a fewer number of PIREPs, as was done in Sharman et al. (2006), an estimate of the standard deviation of the AUC can be obtained. Using 200 subsamples of half the number of available PIREPs gives a standard deviation of the AUC for both WRF ARW and WRF NMM of about 0.006, or < 1% of the AUC. Compared to the 4% difference in AUCs for the 24 h forecasts, the difference is thus significant, and is consistent with the higher performance of the individual turbulence metrics.

5. Effect of terrain data

As mentioned in the overall summary report (memorandum entitled “WRF Rapid Refresh dynamic core evaluation and recommendation” to S. Lord and N. Seaman from S. Benjamin and J. Brown, dated 31 Aug. 2006) the terrain data sets were not exactly the same in the AWR and NMM simulations. Eyeball evaluations of the terrain data differences and “smoothness” of the terrain were shown in the summary report but it is difficult to predict its effect on model performance. In an attempt to remove some of these uncertainties (a) more rigorous evaluations of the terrain were performed, and (b) model turbulence diagnostic performance was evaluated separately east of the Rockies, where terrain influences, especially in the generation of gravity waves and turbulence, should be minimal.

a. Quantitative evaluation of terrain differences

The spatial statistics of the terrain were evaluated using second-order spatial structure functions (i.e., the average of the squared difference in terrain height as a function of spacing or lag) in the East-West direction (x) and the North-South direction (y). Since both models have different grid spacing, the average spacing of the structure functions for each lag were calculated. The square root of the structure functions represents the rms height difference as a function of spatial

separation. The results are shown in Fig. 6. Both models have very similar spatial height statistics for the full domain and the western half of the domain, especially in the east-west direction where the staggered E-grid has less effect. Thus the difference in terrain statistics is probably too small to be an effect on the ARW vs. NMM differences in turbulence forecast model results.

b. Evaluation of turbulence metrics for Eastern U. S.

To try to isolate model performance differences from terrain differences, the turbulence forecast product was rerun using the 24 hr Phase 2 datasets but using only pireps east of 102E longitude for all seasons (giving about 4,000 observations) for verification. The standard performance metric of areas under the ROC curves (AUR) were then recalculated for the top 10 diagnostics and the GTG combination. The results are summarized in Table 3. Note that for the GTG combination of the 10 diagnostics (which was not optimal since the weights of the individual diagnostics were evaluated from the entire conus domain) the (ROC) areas (presented as the mean +/- one standard deviation (sd) of 200 subsamples):

NMM area = 0.785 ± 0.0072
ARW area = 0.806 ± 0.0082

These are to be compared with the stats from the entire domain runs:

NMM area = 0.762 ± 0.0064
ARW area = 0.790 ± 0.0063

So for both models, the overall performance was slightly better over the eastern portion of the U.S., probably due to the inability of the current algorithm suite to correctly diagnose mountain wave turbulence events. As expected, because of the fewer number of observations in the east compared to the entire conus, the uncertainty is a little higher, but the difference is still significant: the NMM AUC + 1 sd is still < ARW AUC - 1 sd. The curves do overlap if 2 sds are considered, whereas the ARW curves were still higher at 2 sds when the entire domain was used.

Another measure is to compare the performance of the individual top 10 diagnostics. Of the 9 that are the same for the NMM and ARW sets, all 9 had higher pod curves for the ARW. The chances of this happening at random are of course fairly small.

6. Summary, issues, and recommendations

a. Summary

- The NMM is definitely “smoother” than ARW
  - Qualitatively by inspection of contours of indices
  - Quantified by effective filtering of structure functions
  - WRF ARW and RUC13 have about the same level of filtering

- Based on null-mog POD areas at upper levels
  - Overall diagnostic performance for the ARW core is better than NMM for both 12 and 24 hour forecasts for almost the entire diagnostic suite
  - The differences are larger for 24 h than 12 h forecasts
  - The GTG combination AUCs are about 2%, 4% higher using the WRF ARW core than using WRF NMM, for the 12, 24 h forecasts respectively.
  - The GTG combination for the 12 h forecast based on the WRF ARW core is about the same as the RUC13
  - The performance of the GTG combination was analyzed using 200 subsets of ½ of the available pireps. The mean of the subsets is larger for ARW than for NMM by at least 3 standard deviations.
  - Most diagnostics are better in both versions of WRF than RUC13, but some diagnostics take advantage of isentropic coordinates so for those the RUC13 performance is superior
  - When comparing the WRF core performances only over the eastern half of the U.S where differences in terrain structures are small, the results are a little more uncertain than for the entire domain, but do show that, consistent with the entire CONUS domain findings, the WRF ARW core performs significantly better than the WRF NMM core

b. Issues

- Initialization errors could be swamping forecast errors
- There is some additional evidence of better statistical behavior of ARW vs NMM in the west. Could this be due to
  - the Arakawa C grid is giving a better representation of gravity wave dispersion than the E grid
  - divergence damping in the WRF NMM
  - both?
- The effect of topography smoothing should be evaluated.
- A quantitative evaluation of the effect small scale features on turbulence forecasts needs to be rigorously evaluated.

c. Recommendations

- From the turbulence forecasting point-of-view the WRF ARW core produces consistently superior performance compared to the WRF NMM. Almost every turbulence diagnostic computed has better performance using the ARW core. The GTG combination performance metric (AUC) is about 4-5% higher using the ARW core. Subset sampling shows this result is statistically significant. The goal of a GTG upgrade is to increases AUC performance by about 5%, so this difference is also significant from an operational point-of-view. **Thus we strongly recommend adoption of the WRF ARW core for the initial RR implementation.**

- The effective filtering of both models should be reevaluated and procedures implemented to reduce the amount of filtering. Using pireps for which the pilot specifically stated the turbulence was mountain wave induced, our research has found that turbulence diagnostics that depend on the small scale variability output from the model, such as EDR, consistently outperform other indirect diagnostics (e.g. Ri). Thus resolution of the small scale features is important for turbulence forecasting.
- In the future higher resolution nests imbedded in the operational WRF-RR model over mountainous regions should be considered, and would provide consistently better turbulence forecasts in the western half of the U. S.

7. References

Ellrod, G. P., and D. L. Knapp, 1992: An objective clear-air turbulence forecasting technique: Verification and operational use. *Wea. Forecasting*, **7**, 150-165.

Frehlich, R., and R. Sharman, 2004: Estimates of turbulence from numerical weather prediction model output with applications to turbulence diagnosis and data assimilation. *Mon. Wea. Rev*., **132**, 2308-2324.

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

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

Table 1. Comparison of AUCs for the top 20 turbulence diagnostics and the GTG combination of the top 10 diagnostics based on RUC13, WRF ARW, and WRF NMM DTC Phase 2 12h forecasts. The number of observations (PIREPs) varied slightly from model to model because of missing model runs, but was about 8,000. The * indicates the 10 diagnostics used in the GTG combination. Uncertainty bounds of one standard deviation derived from subset sampling are provided for the GTG combinations.

Index	AUC		
	RUC13	WRF ARW	WRF NMM
1. GTG combination	0.819±0.0044	0.820±0.0056	0.804±0.0049
2. Frontogenesis fn1	0.795*	0.790*	0.790*
3. Ellrod 1	0.779*	0.782*	0.764*
4. Temp gradient	0.748*	0.721	0.712
5. EDR	0.734*	0.795*	0.769*
6. Sat Ri	0.724*	0.729*	0.729*
7. NCSU1	0.723*	0.742*	0.722*
8. sigma w	0.702*	---	0.700
9. DTF3	0.702*	0.742*	0.740*
10. UBF	0.685*	---	0.728*
11. NVA	0.663*	0.682	---
12. Brown2	0.694	0.752*	0.737*
13. Colson-Panofsky	0.715	0.747*	0.744*
14. Brown1	0.673	0.741*	0.719
15. NGM1	0.691	0.737*	0.722
16. Ellrod 2	0.730	0.724	0.727*
17. VWS	0.728	0.704	0.690
18. Ri	0.713	0.720	0.717
19. DTF5	0.666	0.715	0.707
20. Ri (from TW)	0.695	0.692	0.721
21. NGM2	0.635	0.692	0.673
22. Frontogenesis fn2	0.609	0.687	---
23. ABSIA	0.671	0.686	0.685

Table 2. Comparison of AUCs for the common top 18 turbulence diagnostics and the GTG combination of the top 10 diagnostics derived from the WRF ARW and WRF NMM DTC Phase 2 24h forecasts. The number of observations (PIREPs) varied slightly from model to model because of missing model runs, but was about 7,000. The * indicates the 10 diagnostics used in the GTG combination. Uncertainty bounds of one standard deviation derived from subset sampling are provided for the GTG combinations.

Index	AUC	
	WRF ARW	WRF NMM
1. GTG combination	0.790±0.0063	0.762±0.0064
2. Frontogenesis fn1	0.767*	0.747*
3. Ellrod 1	0.760*	0.740*
4. EDR	0.758*	0.719*
5. Brown2	0.740*	0.716*
6. DTF3	0.729*	0.711*
7. Colson-Panofsky	0.726*	0.704*
8. Ellrod 2	0.707	0.703*
9. Sat Ri	0.722*	0.693*
10. NCSU1	0.720*	0.689*
11. Brown 1	0.716*	0.683
12. NGM1	0.713*	0.705*
13. DTF5	0.702	0.687
14 Ri	0.713	0.685
15. Temp gradient	0.712	0.685
16. Ri (from TW)	0.686	0.687
17. VWS	0.692	0.672
18. UBF	---	0.675
19. NGM2	0.676	---
20. Frontogenesis fn2	0.676	0.672
21. ABSIA	0.673	0.666
22. NVA	---	0.655
23. sigma w	0.670	---

Table 3. Comparison of AUCs for the top 10 turbulence diagnostics and the GTG combination of the top 10 diagnostics derived from the WRF ARW and WRF NMM DTC Phase 2 24h forecasts for the eastern half of the U.S. only (east of 102W). The number of observations (PIREPs) varied slightly from model to model because of missing model runs, but was about 4,000. Uncertainty bounds of one standard deviation derived from subset sampling are provided for the GTG combinations.

Index	AUC	
	WRF ARW	WRF NMM
1. GTG combination	0.806±0.0082	0.785±0.0072
2. Frontogenesis fn1	0.788	0.785
3. Ellrod 1	0.781	0.760
4. EDR	0.751	0.744
5. Brown2	0.767	0.733
6. DTF3	0.770	0.738
7. Colson-Panofsky	0.777	0.750
8. Ellrod 2	0.753	0.730
9. Sat Ri	0.756	0.733
10. NGM1	---	0.724
11. DTF5	0.738	---
12 Ri	0.748	0.726

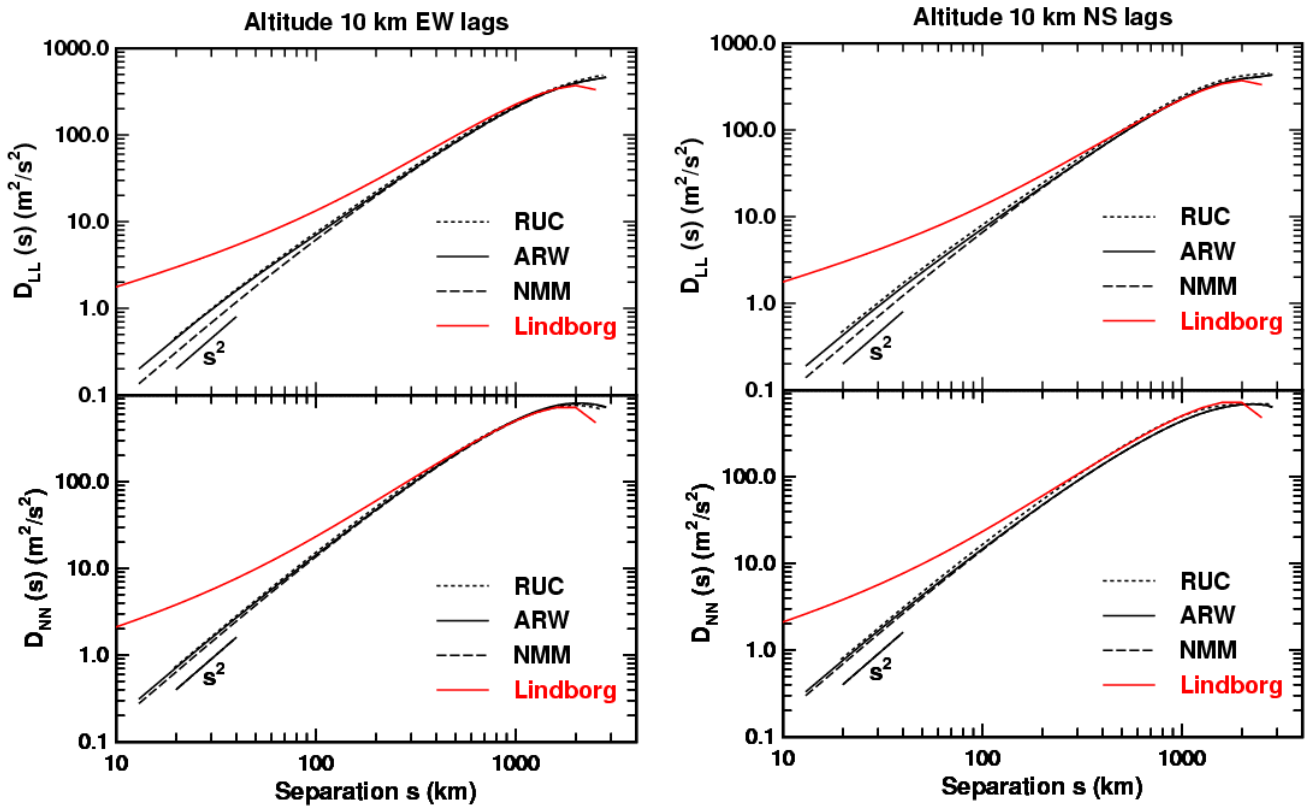

Fig. 1. Example of longitudinal (D_{LL} (s)) and transverse (D_{NN} (s)) structure functions of horizontal velocity at an altitude of 10 km derived from RUC13, WRF ARW and WRF NMM compared with the climatology from aircraft data (red lines) from Lindborg (1999). These results are typical and show slightly more smoothing in the WRF NMM than either the RUC13 or WRF ARW, especially for the longitudinal structure functions.

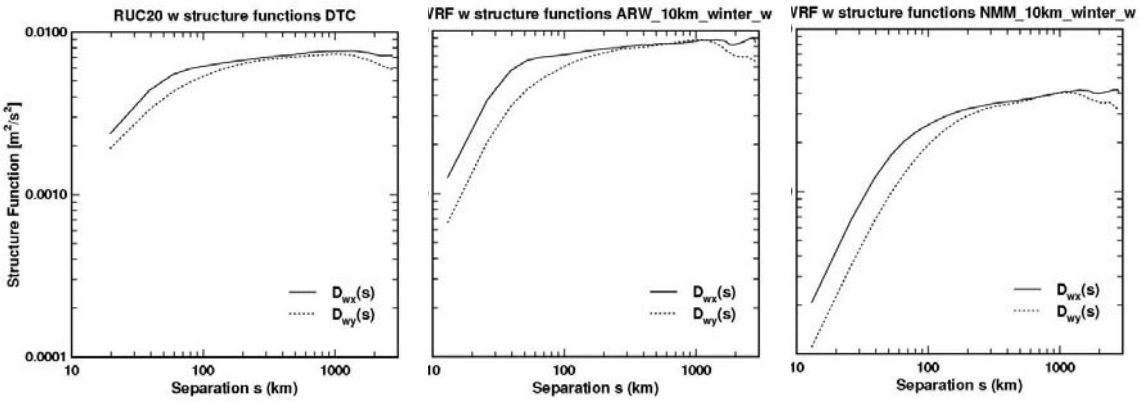

Fig. 2. Example of transverse structure functions of vertical velocity w derived from RUC13 (left panel), WRF ARW (center panel) and WRF NMM (right panel). These results are typical and show a) more small-scale energy in RUC13 compared to both WRFs, b) a consistent level at larger lags between the RUC13 and WRF ARW, and c) lower levels and pronounced filtering at all resolved scales from the WRF NMM compared to either the RUC13 or the WRF ARW.

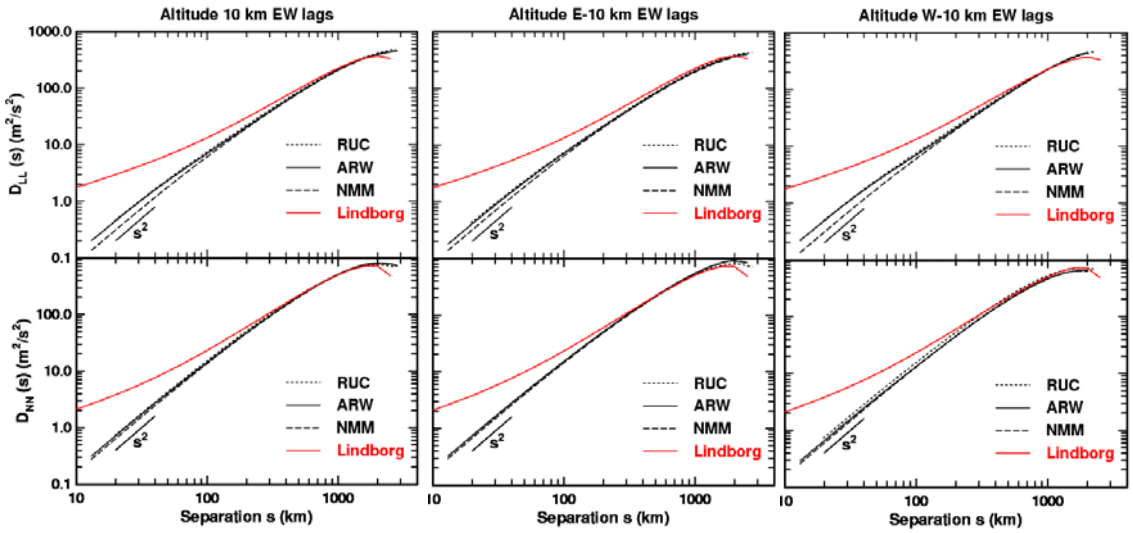

Fig. 3. Example of longitudinal (DLL) and transverse structure (DNN) functions in the east-west direction of horizontal velocity components derived from RUC13, WRF ARW, and WRF NMM, for entire domain (left panel), eastern domain, east of 102 longitude (center panel), and west domain, west of 102 longitude (right panel). The expected result based on aircraft measurements is indicated by the red curve labeled “Lindborg” (Lindborg 1999).

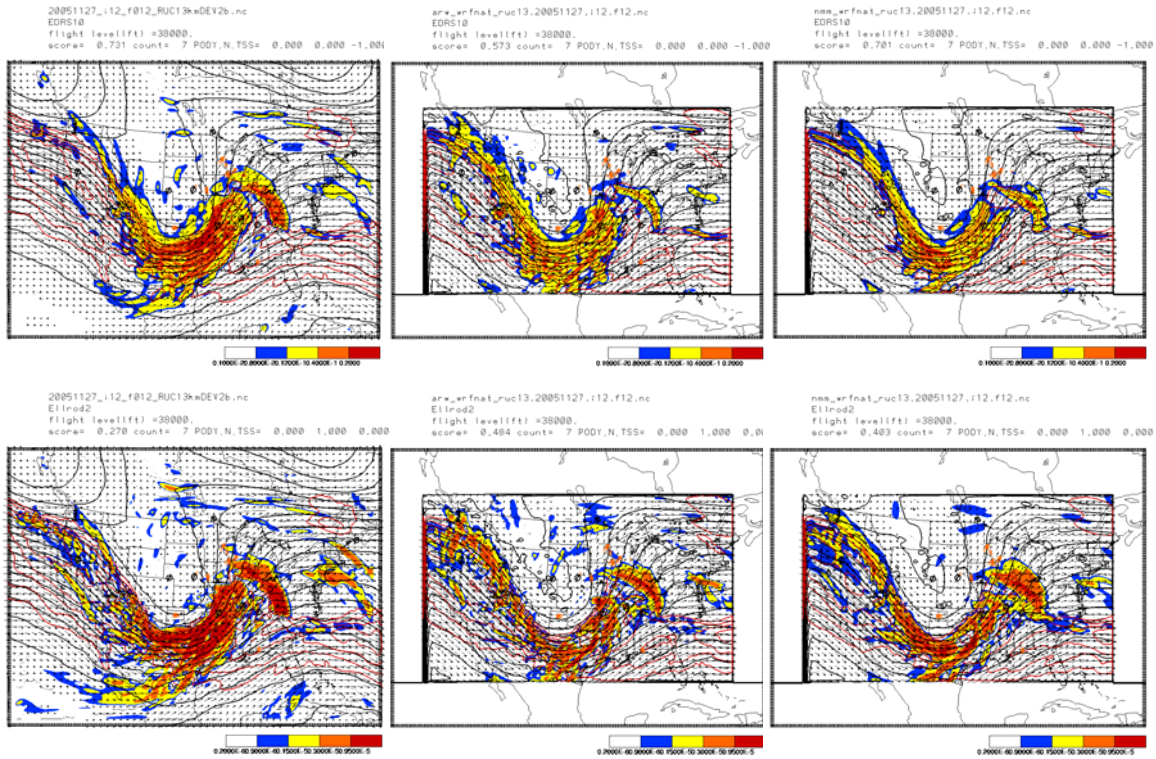

Fig. 4. Contours of Ellrod2 index (Ellrod and Knapp 1992) upper panels and eddy dissipation rate, $\varepsilon^{2/3}$ (Frehlich and Sharman 2004) at FL380 computed from second-order structure functions of horizontal velocity (lower panels) derived from RUC13 (left panels), WRF ARW (center panels), and WRF NMM (right panels) 12 hr forecasts (Phase 1).

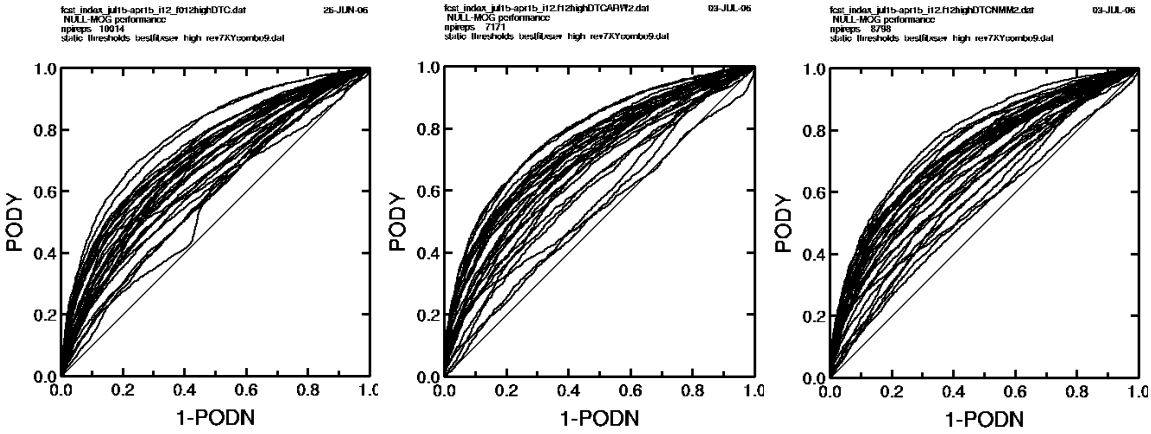

Fig. 5. Null-mog POD curves derived from comparing 12 hr forecasts of 30 turbulence diagnostics derived from RUC13 (left panel), WRF ARW (center panel) and WRF NMM (right panel) for all months at upper levels (Phase 2) to available turbulence PIREPs.

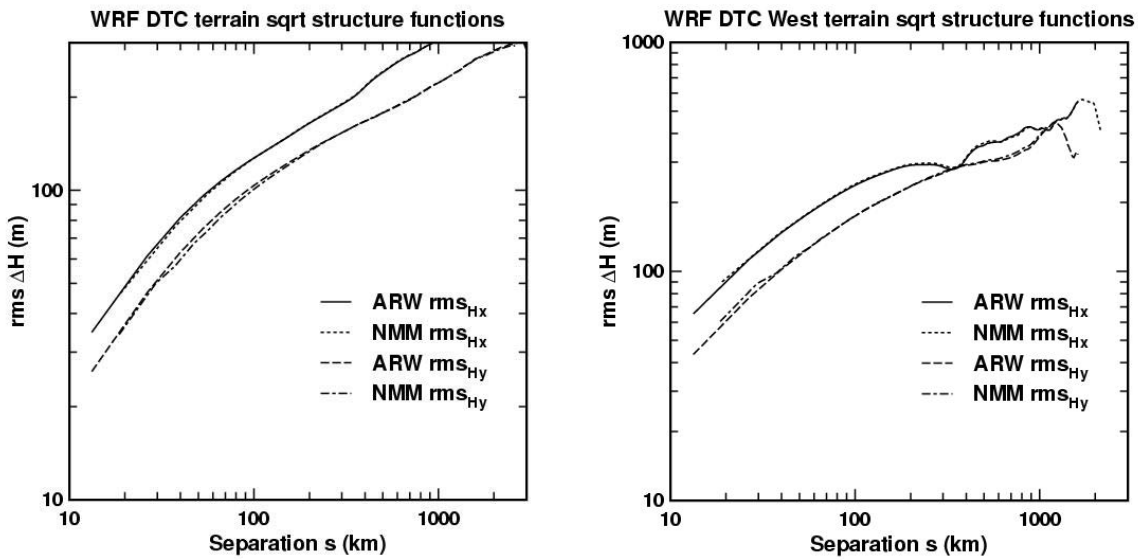

Fig. 6. Comparison of second-order spatial structure functions derived from the WRF ARW and WRF NMM terrain. Upper curves are structure functions in the east-west direction (x) and lower curves are structure functions in the north-south direction (y). Left panel is derived from terrain over the entire conus domain, while the right panel is for terrain west of 102W only.