Major comments:

1) Motivation: What is the objective of the study?

(a) Is it to inform WRF developers of the performance issues?

(b) Is it to tell AQ community what to expect when WRF replaces MM5?

(c) Is it to warn the regulators of potential compliance issues when the transition from MM5 to WRF takes place?

The paper, in its current form, falls short of (a). For example, how much of the differences may be due to "improvements" in WRF mentioned in lines 16-18 on page 1083? Could "mass conservation" be a factor for large differences in performance? Which U* is recommended? What is different in Thompson than Reisner microphysics that may lead to differences in CMAQ?

It comes closest to answering (b) but we do not know if CMAQ was somehow tuned to MM5. A more in depth analysis addressing issues like the following may be helpful. If higher O3 with WRF is due to less dry deposition, doesn't the same (i.e., less dry deposition) also apply to Midwest? Can you quantify, or at least show, the effect of dry deposition. An increase of 16 ppb in summertime mean bias, even if limited to some sites along the Gulf Coast, is alarming. Can you run a sensitivity to cloud fraction?

The paper also does not respond to (c). The regulators would want to know how difficult/easy it would be to show compliance with NAAQS whenMM5 is replaced with WRF.

The paper may be more interesting if it is focused on one of these questions.

2) Analysis: What is the underlying assumption of Equations (3) and (4)? That the bias/error has the same distribution as the observations? What if the distributions are skewed differently? In my opinion, the normalization should be carried on each individual observation. Figures 3, and 5 show the monthly average differences. Give us a sense of the spatial distribution of monthly average concentrations! Where are the differences more significant relative to concentrations? From Table 2 and Equation 3, I found that the median hourly O3 is 23 ppb. A difference of 2 ppb in monthly average ozone means that the differences in hourly ozone may be significant. How is the diurnal distribution of the differences? Same for PM2.5. I found that the median SO4 is 1 ug/m3 (IMPROVE) so 1 ug/m3 difference in monthly average may be substantial. As for NO3, I find the median to be 0.35 (IMPROVE) so a difference of 1 ug/m3 is definitely not "small". There is a mismatch in time scales. While meteorological performance (Figure 1) is daily and hourly, chemical performance is monthly average (Figs. 3&5). I have no way of accessing the meteorological model performance assessment in Gilliam and Pleim (2009).

3) Method: "appear to be" on line 13 page 1095: How do you see? How do you identify the most important factors (line 18 page 1096)? For example, on line 13 page 1097, you mention the use of a tool (the sulfur tracking version of CMAQ). How about elsewhere? Did you confirm that OH concentrations are higher along the Gulf of Mexico (lines 27-28 page 1097). Recently, I heard that there may be serious problems with CMAQ's OH concentrations in the upper troposphere.

Explain the method you used for identifying the reasons. Did you use process analysis? Did you run sensitivity tests? Do you have quantitative evidence that (what you call) "major factors" are really major factors?

4) Language: Given that you probably did not use a method that allows you to quantitatively compare the impact of different factors, firm language such as “were caused by” (line 14 page 1100), “due to less dry deposition of HNO3” (line 21 page 1100), “primarily due to” (line 13 page 1101) is inappropriate. More cautious language such as “likely plays a role” (line 7-8 page 1096), “are likely related” (line 18 page 1100), “appears to be most directly related” (line 1 page 1101) is acceptable.

Minor Comments:

1) Line 18 page 1083: Need a reference(s)
2) Line 17 page 1085: it is not necessary to state that layer collapsing is not used.
3) I believe the time-of-the-day references (i.e., morning, afternoon, night) in meteorological model performance assessment of Section 3 are based on UTC. This is confusing because local time is 5-6 hours behind UTC.
4) Lines 17-18 page 1090: it is not clear why you want to make WRF parameterization more like MM5’s. This statement is probably unnecessary.
5) Line 3 page 1091: Is there any data to compare to predicted cloud fractions? Is it a good thing that WRF has greater cloud fraction?
6) Line 8 page 1092: Similarly, is there any dry deposition data to compare to? Is greater dry deposition a step in the right direction?
7) Lines 8-26 on page 1093. Is this discussion (how PM2.5 mass is calculated) necessary?
8) Line 16 page 1094: explain why this was “expected”
9) Line 3 page 1095: the correct reference is to Table 3. Same thing in line 1 and 8 of page 1097
10) Line 5 page 1095: you are mixing 1-hr O3 with 8-hr O3.
11) Line 25 page 1095: I suppose these times are local and not UTC (See minor comment #3 above).
12) Lines 1-4 page 1096: Were all these areas VOC limited?
13) Section 4.2.3: Why is TNO3 so much higher then NO3 in WRF-CMAQ along the Gulf Coast?
14) Lines 20-21 page 1098: By “the same factors that result in the higher O3” do you also mean the gas-phase? If so, how? There is probably something more basic such as a lower mixing height (along the Gulf Coast).
15) Line 2 page 1100 “requires further investigation”: There are no large differences in precipitation, plus SO4 was much lower, but SO4 wet deposition is higher. How confident are you with this result?
16) Figure 1: The legends are only visible at 4x magnification.
17) Figure 3: The scales are hardly visible at 2x magnification. The caption should define the difference as WRF minus MM5.