Geosci. Model Dev. Discuss., 6, C1239–C1262, 2013
www.geosci-model-dev-discuss.net/6/C1239/2013/

Geoscientific
Model Development
Discussions
Open Access

Interactive comment on "Evaluation of the United States National Air Quality Forecast Capability experimental real-time predictions in 2010 using Air Quality System ozone and NO_2 measurements" *by* T. Chai et al.

T. Chai et al.

Tianfeng.Chai@NOAA.gov

Received and published: 16 August 2013

We are very grateful to the reviewers for reading the manuscript extremely carefully and forwarding their valuable suggestions for improvement. Point-by-point responses to the reviewers' comments are listed below.

Responses to the comments of referee #1:

General Comments:

My main comment on the manuscript is the intended goal, which is not too clear from the beginning. The best description of the goal might be present in the very last paragraph, which starts to say that "The type of analysis presented here has guided recent updates". This is also mentioned as last sentence in the abstract, but throughout the paper I hardly find it back. Another goal might be that the paper will serve as a reference for future evaluations, or in papers focusing on the model improvements; if this is the case, it should be addressed more clear. Will this type of analysis be done every year, with some overview after 10 years or so? The paper also mentions at some occasions the difference between the operational system and the experimental system; results are solely for the experimental system. Isn't one of the goals that the experimental system in the end should replace the current operational system, when it is proven to have similar or preferably better skills ? This would require a more detailed comparison between the skills of the two systems, which is probably left for other publications; if this is planned, it should be mentioned.

Providing a reference for future evaluations and model developments is the main objective for the paper. This has been clarified in Introduction.

Although the evaluation results have been used to guide recent updates, description on the detail is beyond the scope of our current paper. The last sentence of the Abstract, “Based on the analysis presented here, experimental ozone prediction system was updated for summer 2012” has been deleted.

The authors share the foresight of the reviewer that periodic evaluation documentation of the prediction systems is highly desirable - especially when there occurred major upgrades of the systems (e.g., Eder et al. 2006 and 2009, Gorline and Lee 2009a and 2009b).

As the reviewer pointed out, a detailed comparison between the skills of the operational and the experimental systems would require a separate publication. However, Saylor and Stein(2012) provided some details on this topic. It was cited in this paper to explain the evaluation results presented by Gorline and Lee(2009b).

Gorline, J. L. and Lee, P.: Performance of NOAA-EPA air quality predictions, 2007 - 2009, in: the 8th Annual CMAS Conference, Chapel Hill, NC, October 19-21, 2009b.

Saylor, R. D. and Stein, A. F.: Identifying the causes of differences in ozone production from the CB05 and CBMIV chemical mechanisms, Geosci. Model Dev., 5, 257–268, doi:10.5194/gmd-5-257-2012, 2012.

Specific Comments:

- *Section 3*
  How many sites are actually used, and how distributed over rural and (sub)urban

  classes? Might have missed that while reading. From figures 8-9 one can see that some regions have hardly any stations; could this have an effect on the conclusions ?

  1124 ozone and 408 NO_2 sites are actually used. Among 1124 ozone sites, there are 200 urban, 455 suburban, 462 rural, and 7 unknown. The number of NO_2 monitors at urban, suburban, rural, and unknown settings are 130, 148, 126, and 4, respectively. They have been added to text.

  We agree that the sparseness of AQS stations in some regions, especially for NO_2 monitoring could affect the conclusions. The following statement has been added to the discussion section. “There are several limitations in our current evaluation study. For instance, the AQS stations are quite sparse in some regions, especially for NO_2 monitoring.”

- *Section 3 and Figure 2*
  Why could be (in summary) the reason for differences? The time shift is mentioned, but what about other causes (other correction factors, conversion to standards, etc)? The use of circles in Figure 2 makes the spread much bigger than it is, given the very high correlation coefficient. A density plot would be more useful here. Correlation coefficients could be put in the figure too.

  Other than the time shift, another reason for the differences in Figure 3a is the quality control work carried out after AIRNow reporting. This has been mentioned in text. The quality control might be related to correction factors, conversion to standards, etc.

Scatter plots in Figure 3 have been replaced by density plots. As suggested, R^2 values have been put in the figure too.

- *Section 3*
  Very elegant method of testing on time shifts that might be present. This takes a large part of the section, might be useful to put this in a subsection

  Section 3 has been divided into three subsections. Testing on time shifts is now put in Section 3.2 “Consistency check of ozone observations”.

- *Section 3, p 2619, line 6*
  What is meant with “Without NO2 measurements from AIRNow for the extra checking ...” . Isn’t only AQS used for the evaluation?

  Yes, only AQS NO_2 measurements are used for the evaluation. There are no AIRNow NO_2 data available to perform the similar consistency examinations between AIRNow and AQS as what is done for ozone. This has been explained in text.

- *Section 4, p 2619, line 13-14*
  Is it possible to explain shortly the observation representation methods referred to?

  The “direct matching” referred here means that there is no interpolation applied to model results. It is consistent with the previous NAQFC evaluation studies. However, there is a slight difference from what is described in Eder et al (2009), where multiple observations inside a single grid cell are averaged as the rep-

  resentative measurement for the grid cell. In this paper, each measurement is compared against model prediction independently when there are two or more monitors located within one grid cell. The explanation has been added.

- *Section 4.1*
  For some types of NO2 monitors it is suggested that these actually observe NOy, see for example www.atmos-chem-phys.net/7/2691/2007/ . Is this taken into account? Might increase the bias even more, but will help to judge how well the model represents the observations.

  We used AQS NO_2 measurements without any correction. This issue has been mentioned and the following two references have been added.

  Dunlea, E. J., Herndon, S. C., Nelson, D. D., Volkamer, R. M., San Martini, F., Sheehy, P. M., Zahniser, M. S., Shorter, J. H., Wormhoudt, J. C., Lamb, B. K., Allwine, E. J., Gaffney, J. S., Marley, N. A., Grutter, M., Marquez, C., Blanco, S., Cardenas, B., Retama, A., Ramos Villegas, C. R., Kolb, C. E., Molina, L. T., and Molina, M. J.: Evaluation of nitrogen dioxide chemiluminescence monitors in a polluted urban environment, Atmos. Chem. Phys., 7, 2691–2704, doi:10.5194/acp-7-2691-2007,

  Steinbacher, M., Zellweger, C., Schwarzenbach, B., Bugmann, S., Buchmann, B., Ordóñez, C., Prevot, A. S. H., and Hueglin, C.: Nitrogen oxide measurements at rural sites in Switzerland: Bias of conventional measurement techniques, J. of Geophys. Res., 112, D11 307, doi:10.1029/2006JD007971, 2007.

- *Section 4.1*

Pictures instead of tables might help to quickly interpretted the statistics. This holds for tables 1-4 and 7-8.

We agree that figures help to quickly interpret the statistics. For documentation purpose, we feel that tables have the advantage to precisely present the results. As we have many figures already, we chose to put some results in tables when feasible.

- *Section 4.2*
  Why showing pictures for 2 summer months? Differences are not large.

  Figures for July have been removed. Figs. 8 and 9 have been combined into one figure. Text has been modified to reflect the change.

- *Section 4.3*
  The acronyms HIT, CSI, etc are only expanded in the introduction; might be useful to have them here too. Also an intuitive explanation of what each measure tells you would be be useful, not this is only done for ETS.

  The acronyms HIT, CSI, etc are expanded in this section. Brief descriptions of them are added to text.

- *Section 4.5, p 2626, line 9*
  Is local time (LT) including the daylight saving regime, or is it actually a local standard time based on the longitude?

  The local time here does not include daylight saving regime and it is based on the official time zone instead of the longitude. The time zone of each AQS site is

provided in the site description file. This has been clarified in text.

Technical corrections:

- *p 2616, line 2*
  Isn't the second 48 hour forecast produced for 00Z instead of 06Z?

  The second 48 hour forecast is produced for 06Z in order to disseminate results at convenient times. Current 06Z and 12Z forecast products are available in the early morning and early afternoon eastern time, respectively.

- *Fig 13, caption*
  Figure does not show the bias, but just the concentrations I guess.

  Corrected.

Responses to the comments of referee #2:

1. General comments:

- *The other aim of the paper is mentioned as "a view towards" the improvement of the model (see, e.g. the abstract). However, the discussion and interpretation of the results is very limited, and for this aspect the authors basically refer to the preprint of Stajner et al. 2013. To my opinion the line may be removed from* *the abstract (line 9), limiting the aim of the paper to the documentation of the performance of the NAQFC system against routine air quality observations.*

We agree that the main objective of the paper is to document the performance of the NAQFC system against routine air quality observations. As suggested, “with a view towards their improvement” has been removed. The last sentence in the abstract, “Based on the analysis presented here, experimental ozone prediction system was updated for summer 2012”, is also deleted for the same reason.

- *The paper evaluates only the first 24 hours of the forecast. Why this limitation? It would be valuable to have at least one figure added which discusses the performance of the second forecast day compared to the first day. I suggest that such a result is added before publishing.*

  Our performance evaluations for previous years demonstrate that there is not much difference between results of the first 24 hours and those of the second 24 hours. A recent publication in GMD by Savage et al. (2013) also demonstrated that there is little difference between day 1 and day 2 ozone forecasts for all metrics using Met Office Unified Model results over the period May 2010 to April 2011. We have cited the paper in text.

  Savage, N. H., Agnew, P., Davis, L. S., Ordóñez, C., Thorpe, R., Johnson, C. E., O’Connor, F. M., and Dalvi, M.: Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial evaluation, Geosci. Model Dev., 6, 353–372, doi:10.5194/gmd-6-353-2013, 2013.

  The differences between forecasts with different lead times are caused by the

change of meteorological fields, but the CTM prediction errors mostly come from the other factors, such as emission sources, chemical mechanism, top and lateral boundary conditions. It is definitely an interesting topic to discuss such differences. However, we feel that it is beyond the scope of the paper.

- *In the abstract I miss references to similar forecast systems worldwide (Europe, Asia). For instance the European model air quality forecast ensemble (gmes-atmosphere.eu) is an interesting comparable capability which should be referred to and the performance may be compared. References should be added, as well as some remarks on the comparisons.*

  The following paragraph and references are added.

  Many operational air quality forecasting systems using 3D CTMs exist worldwide. In Europe, atmospheric composition forecast products have been delivered under the Monitoring Atmospheric Composition and Climate-Interim Implementation project as part of the pre-operational GMES Atmosphere Service (http://www.gmes-atmosphere.eu, see also Menut and Bessagnet, 2010). Similar forecasts are also available in Japan (Maki, 2012), Taiwan(http://taqm.epa.gov.tw/taqm/), and Canada (Talbot et al., 2008). Zhang et al. (2012) summarized some recent real-time air quality forecasting evaluation results. Among all evaluation statistics for hourly ozone, the median positive MB, negative MB, and RMSE, are +4.5 ppbv, -8.1 ppbv, and 16.8 ppbv, respectively. For daily maximum eight-hour ozone categorical statistics, the median POC, CSI, TS, and HIT, are 0.92, 0.18, 0.32, and 0.65, respectively.

- *To my opinion the text can be shortened in several places. The paper is written carefully and provides all the details needed to understand what is done. However, the text in section 4 has several long discussions listing the details in the figures. The figures and tables contain this information already and to my opinion the text should only highlight and summarize the main features (and not the details). I would suggest that the authors look where the text may be shortened.*

  We looked carefully and shortened text at several places. For instance, "For the Pacific Coast, Lower Middle, and Southeast regions, the model biases from Tuesday to Thursday are lower than those on the other days. In the Rocky Mountain region, model biases are lower from Monday to Wednesday. In the Upper Middle and Northeast regions, the model biases on Friday are significantly lower than those on the other days" on page 2615 (lines 14-18) is replaced by a single sentence "Over CONUS and most regions, O_3 biases are higher on weekends than on weekdays."

  In addition, Figs.8 and 9 have been combined by keeping August results only. Almost half of the rows are removed from Tables 5 and 6. Text has been shortened in reflecting the changes.

- *I was a bit surprised the authors limit the statistics metrics to rather traditional bias, RMSE and exceedance scores. In particular, RMSE is well-known to be sensitive to outliers. Furthermore, when the bias is large, RMSE is not an independent measure. Please motivate the choice of metrics.*

  We calculated mean-absolute errors (MAEs) before and found that the basic

trend is same, in spite of reduced values in MAEs than in RMSEs.

In general, RMSEs are suitable to describe variables that are expected to have a Gaussian distribution, and MAEs are for variables that have a uniform distribution. The distribution of model errors is clearly a Gaussian distribution, but with bias. Practically, the key difference between the two metrics is whether to penalize the large errors or to give them the same weight as other small errors. We believe it is reasonable to penalize the large errors as we want to identify the model problems that cause these.

We understand that the RMSE can be greatly affected by the model bias, but it is still an independent statistical measure unless it is dominated by a couple of outliers. The dominance of RMSE by several outliers can happen if extreme large errors exist. In the current case, there are no such extreme model-observation values for either ozone or NO_2.

Probably there is still no consensus on choosing RMSEs or MAE. Recent summary by Zhang, Y. et al (2012) show that most of the forecasting evaluations provide RMSE values.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., and Baklanov, A.: Real-time air quality forecasting, part I: History, techniques, and current status, 60, 632–655, doi:10.1016/j.atmosenv.2012.06.031, 2012

When addressing this question, we found serious flaws in the two papers (shown below) that promote the MAE over the RMSE. We are preparing a technical note to address issues relating to choosing between the MAE and the RMSE. The technical note will be submitted to *Atmospheric Environment soon.*

Willmott, C., Matsuura, K., DEC 19 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82.

Willmott, C. J., Matsuura, K., Robeson, S. M., JAN 2009. Ambiguities inherent in sums-of-squares-based error statistics. Atmos. Environ. 43 (3), 749–752.

- *The paper of Stajner, 2013 is not available to me. This seems to contain a lot of the interpretation of the results presented here. The authors should make sure that the overlap is minimal.*

  The paper of Stajner, 2013 focuses on recent updates since 2010. However, it has not been accepted for publication yet and thus it has been removed from references.

- *Sensors with molybdenum converters suffer from an overestimate of NO2. From the paper it is not clear to me if corrections were applied to the NO2 data to account for this issue.*

  In this study, the AQS NO_2 measurements were used without any correction. This has been clarified in Section 3.

- *Why is especially the SouthEast difficult to model correctly ?*

  Our results show that NO_2 are overestimated in SouthEast, especially in summer. The region is mostly in NO_x-sensitive chemical regimes during the summer, as it has abundant VOC sources from forests but does not have as much NO_2 sources

as the other regions. Thus the overestimated NO_2 can produce ozone much more efficiently than in the other regions. The discussion is included in Section 4.1.

2. Detailed comments:

2.1 Section 2

- *Please discuss also the boundary conditions? Biases in ozone could partly be caused by long-range transport and influx from the stratosphere. What is known from previous work on the quality of these aspects of the model?*

  Lateral boundaries are described and discussed. The following two references have been added when discussing long-range transport and influx from the stratosphere.

  Browell, E. V., Hair, J. W., Butler, C. F., Grant, W. B., DeYoung, R. J., Fenn, M. A., Brackett, V. G., Clayton, M. B., Brasseur, L. A., Harper, D. B., Ridley, B. A., Klonecki, A. A., Hess, P. G., Emmons, L. K., Tie, X., Atlas, E. L., Cantrell, C. A., Wimmers, A. J., Blake, D. R., Coffey, M. T., Hannigan, J. W., Dibb, J. E., Talbot, R. W., Flocke, F., Weinheimer, A. J., Fried, A., Wert, B., Snow, J. A., and Lefer, B. L.: Ozone, aerosol, potential vorticity, and trace gas trends observed at high-latitudes over North America from February to May 2000, J. of Geophys. Res., 108, doi:10.1029/2001JD001390, 2003.

  Tang, Y., Lee, P., Tsidulko, M., Huang, H.-C., McQueen, J., DiMego, G., Emmons, L., Pierce, R., Thompson, A., Lin, H.-M., Kang, D., Tong, D., Yu, S., Mathur, R., Pleim, J., Otte, T., Pouliot, G., Young, J., Schere, K., Davidson, P., and Stajner, I.: The impact of chemical lateral boundary conditions on CMAQ predictions of tropospheric ozone over the continental United States, Environ. Fluid Mech., 9, 43–58, doi:10.1007/s10652-008-9092-5, 2009.

- *Wildfire emissions: these are based on a multi-year averaged emission. Wildfires are an important source of variability and there are several efforts worldwide to use space observations to improve estimates on a monthly/weekly/daily time scale, also in near- real time. Please comment.*

  Comments on ignoring the temporal and spatial variability of the wild fire emission sources are included. Current effort of incorporating NOAA Smoke Forecasting System (SFS) to provide the CMAQ model with near-real-time emissions from large wildfire and agricultural burning is also mentioned. Five related references are added.

- *What about the 100 hPa "zero flux assumption" (p2615): does this lead to a reasonable ozone concentration in the upper troposphere? Is there previous work studying the free troposphere ozone concentrations of the system?*

  We see reasonable ozone concentration results in the upper troposphere. However, not considering the stratospheric ozone intrusion may cause ozone underestimations at high-latitudes. This limitation has been stated.

  There is previous work studying the free troposphere ozone concentrations of the current NAQFC system (see references below). The references are not included

in the paper as the method has not been implemented.

Mathur, R., Lin, H. M., McKeen, S., Kang, D., and Wong, D.: Three-dimensional model studies of exchange processes in the troposphere: use of potential vorticity to specify aloft O_3 in regional models, The 7th Annual CMAS Conference, Chapel Hill, NC, USA, 2008

Lin, H., R. Mathur, S. A. McKeen, P. C. Lee, and J. Mcqueen: Application of Potential Vorticity in a comprehensive air quality forecast model for Ozone, 89th American Meteorological Society Annual Meeting, New Orleans, LA., 20-24 Jan., 2008.

2.2 Section 3

- *Near-real time data: Please expand the discussion on how the near-real time data compare to the quality controlled data sets which become available after some time. Are there many outliers?*

  Section 3.2 "Consistency check of ozone observations" is devoted to discuss the comparison between AQS and AIRNow data. New density plots are now given in Fig.3. Fig.3a provides a snapshot on 31 May, 2010. As shown in Fig.3a, there are outliers, but the majority falls to 1:1 line, as expected. After removing the stations suspected for the time-shifting problem, we see less outliers and better matching between AQS and AIRNow data.

- *p17 l15: "valid hourly observations": What does "valid" mean?*

There are quality flags associated with each AIRNow observation data entry. Only “G" (meaning good) flagged data are used here. As most AIRNow data users are aware of this, “valid” is removed to avoid confusion.

- *p17 l19: 10000 observations: please also mention how many stations are involved.*

  There are 1124 ozone and 408 NO_2 AQS sites actually used in the paper. Among 1124 ozone sites, there are 200 urban, 455 suburban, 462 rural, and 7 unknown. The number of NO_2 monitors at urban, suburban, rural, and unknown settings are 130, 148, 126, and 4, respectively. They have been added to text.

- *Fig 2: the term “overlapped” is not clear: is it the number of overlapping observations, or the total after removal of the overlap. If observations overlap, is one of them kept? What is “overlap*” ? This should be explained also in the caption of the figure.*

  In Fig 2, “O_3 overlap” is the number of overlapping ozone observation pairs and “O_3 overlap*” is the number of observation pairs after removing measurements from the 74 questionable sites. The caption of Fig 2 has been modified. Change is also made in text to clarify this.

- *p18 l18: in-line formula: why is there a division by 10 ?*

  Here the factor of 0.1, i.e. $\frac{1}{10}$, results from trial and error. Other values, including 0.01 and 0.001, have been tried. By plotting the time series of two AQS-AIRNow differences (with and without time shift) and checking those L2-norm values, 0.1

is found to be a good threshold value to identify time shift problems.

- *Section 3 is a bit long and may be condensed. The overlap and time shift issues are details.*

  We believe that the details for the overlap and time shift issues will be useful to the future AQS data users. Suggested by the other reviewer, this section has been organized into several subsections.

- *I was wondering about the NO2 observations. Sensors with molybdenum converters suffer from an overestimate of NO2, see e.g. Steinbacher et al, doi 10.1029/2006JD007971. Were corrections applied? Please discuss this.*

  In this study, the AQS NO_2 measurements were used without correction. This issue has been discussed and the following two references have been added.

  Dunlea, E. J., Herndon, S. C., Nelson, D. D., Volkamer, R. M., San Martini, F., Sheehy, P. M., Zahniser, M. S., Shorter, J. H., Wormhoudt, J. C., Lamb, B. K., Allwine, E. J., Gaffney, J. S., Marley, N. A., Grutter, M., Marquez, C., Blanco, S., Cardenas, B., Retama, A., Ramos Villegas, C. R., Kolb, C. E., Molina, L. T., and Molina, M. J.: Evaluation of nitrogen dioxide chemiluminescence monitors in a polluted urban environment, Atmos. Chem. Phys., 7, 2691–2704, doi:10.5194/acp-7-2691-2007,

  Steinbacher, M., Zellweger, C., Schwarzenbach, B., Bugmann, S., Buchmann, B., Ordóñez, C., Prevot, A. S. H., and Hueglin, C.: Nitrogen oxide measurements at rural sites in Switzerland: Bias of conventional measurement techniques, J. of Geophys. Res., 112, D11 307, doi:10.1029/2006JD007971, 2007.

2.3 Section 4

- *Why do authors use RMSE: this is sensitive to outliers. The mean-absolute difference would be a good alternative!? RMSE has large contribution coming from the bias, so it is not an independent statistical measure.*

  Note the same question is raised in "General comments" part. Answer below is copied from above.

  We calculated mean-absolute errors (MAEs) before and found that the basic trend is same, in spite of reduced values in MAEs than in RMSEs.

  In general, RMSEs are suitable to describe variables that are expected to have a Gaussian distribution, and MAEs are for variables that have a uniform distribution. The distribution of model errors is clearly a Gaussian distribution, but with bias. Practically, the key difference between the two metrics is whether to penalize the large errors or to give them the same weight as other small errors. We believe it is reasonable to penalize the large errors as we want to identify the model problems that cause these.

  We understand that the RMSE can be greatly affected by the model bias, but it is still an independent statistical measure unless it is dominated by a couple of outliers. The dominance of RMSE by several outliers can happen if extreme large errors exist. In the current case, there are no such extreme model-observation values for either ozone or NO_2.

  Probably there is still no consensus on choosing RMSEs or MAE. Recent summary by Zhang, Y. et al (2012) show that most of the forecasting evaluations

  provide RMSE values.

  Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., and Baklanov, A.: Real-time air quality forecasting, part I: History, techniques, and current status, 60, 632–655, doi:10.1016/j.atmosenv.2012.06.031, 2012

  When addressing this question, we found serious flaws in the two papers (shown below) that promote the MAE over the RMSE. We are preparing a technical note to address issues relating to choosing between the MAE and the RMSE. The technical note will be submitted to *Atmospheric Environment soon.*

  Willmott, C., Matsuura, K., DEC 19 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82.

  Willmott, C. J., Matsuura, K., Robeson, S. M., JAN 2009. Ambiguities inherent in sums-of-squares-based error statistics. Atmos. Environ. 43 (3), 749–752.

- *Why are there different panels for July and August ? If the conclusions are similar these may be combined. For me it would be more useful to see Summer JJA and winter DJF results.*

  Figures for July have been removed. Figs. 8 and 9 are now combined into one figure. Text has been modified to reflect the change. Although adding winter results could be useful, they are not included here considering the length of the paper.

- *p21, l4: It is not surprising the Rocky mountains have the smallest bias because*

NOx is low in this region. Please quote also the relative errors compared to e.g. monthly and regional mean. In general, throughout the paper, it would be useful to have errors also as relative numbers.

The relative NO_2 bias in Rocky mountain is still the smallest, with its annual mean of 4.2%. Its monthly average relative biases range from -7.1% in January to +33.0% in July. We have quoted the normalized mean bias values to text. Additional such relative numbers and brief descriptions are included as well.

As the mean values for CONUS and all regions have been plotted in Fig. 4-7, the relative number are easy to calculate from those MBs and RMSEs. Nonetheless, some of the relative errors have been explicitly presented (e.g., see lines 1-4 on page 2620 in the previous version).

- *p22: The exceedance scores are influenced considerably by the large bias. In Fig. 10 "d" is close to zero. Please discuss this point.*

  We have added discussion on this point and included the following reference.

  Kang, D., Mathur, R., and Rao, S. T.: Real-time bias-adjusted O_3 and PM2.5 air quality index forecasts and their performance evaluations over the continental United States, 44, 2203–2212, doi:10.1016/j.atmosenv.2010.03.017, 2010.

- *There is a lot of tables and a lot of numbers provided. This is in part useful (e.g. to document the results for the different regions) but sometimes similar results are shown in different tables. In particular, I would suggest to remove table 5 (Summer only is enough because that is when exceedances occur). In fig.6*

  it may be considered to remove one of the two limits (70 or 75) because they are close together and the results are similar. Alternative: for 70 keep only the CONUS row.

  We take the suggestion of keeping only the CONUS row for 70 ppbv threshold in Tables 5 and 6. Text has been modified accordingly.

2.4 Section 5

- *p28 l14 "... in order to expose systematic model errors, which could be corrected in the future to improve NAQFC predictions" There is little interpretation of the results. Is it really the emissions which are to blame? Could it be a lifetime issue as well. What about free troposphere, influx from stratosphere ?*

  The emissions cannot be the only reason for the model errors. Based on our results, we just speculate that emissions might be the most important factor. Text has been modified to avoid confusion. In addition, we added the following sentence to the end of this paragraph.

  "It should be noted that other factors, such as chemical mechanism, not considering long range transport at lateral boundaries or ozone intrusion from stratosphere at domain top, all contribute to the current model errors."

- *p29, l8 "monthly mean profiles from global model simulations for most chemical species" Which model is used? Has this been validated? Provide a reference please.*

It is GEOS-CHEM model. We compared the results before and after using the monthly mean profiles and found improvements over constant profiles in the predictions. The following reference on GEOS-CHEM has been added.

Bey, I., Jacob, D., Yantosca, R., Logan, J., Field, B., Fiore, A., Li, Q., Liu, H., Mickley, L., and Schultz, M.: Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation, J. of Geophys. Res., 106, 23 073–23 095, 2001.

- *l9: "Dry deposition was modified based on the Monin-Obukhov similarity theory as well as by including canopy height and density based on recent satellite observations" please provide a reference.*

  The following two references have been added.

  Wu, Y., Brashers, B., Finkelstein, P. L., and Pleim, J. E.: A multilayer biochemical dry deposition model. 1. Model formulation, J. of Geophys. Res., 108, doi:10.1029/2002JD002293, 2003.

  Lefsky, M. A.: A global forest canopy height map from the Moderate Resolution Imaging Spectroradiometer and the Geoscience Laser Altimeter System, Geophysical Research Letters, 37, doi:10.1029/2010GL043622, 2010.

- *l11 "Planetary boundary layer height was constrained to be at least 50 m." How and when does this influence the results.*

  This mitigated the previous high ozone bias problem due to low PBLs at areas close to large water bodies. It also allows dilution of the mobile emissions near

  urban centers and lessened the severity of ozone titration at nighttime. This has been added to text.

- *l14 "The emission data sets have been updated in June 2012, with a pronounced decrease in mobile NOx emissions." By how much?*

  By about 35%. Text has been modified.