Dynamic Evaluation of CMAQ Part II: Evaluation of relative response factor metrics for ozone attainment demonstrations

Kristen M. Foley*, Patrick Dolwick, Christian Hogrefe, Heather Simon, Brian Timin, Norm Possiel

United States Environmental Protection Agency, Research Triangle Park, NC 27711, USA

*Corresponding Author: telephone (919) 541-5367; fax (919) 541-1379; email foley.kristen@epa.gov.

Abstract

The U.S. Environmental Protection Agency provides guidelines on the use of air quality models for projecting whether an emission reduction strategy will lead to future pollutant levels that are at or below the National Ambient Air Quality Standards (NAAQS). The EPA's guidance document for ozone attainment demonstrations recommends an attainment test for the 8-hour ozone NAAQS based on using the ratio of output from "future" and "base" model simulations through the calculation of location-specific Relative Response Factors (RRF). The 2007 guidance document as well as other related studies have recommended the use of retrospective evaluation studies in order to evaluate the ability of an air quality model to represent a change in air quality (dynamic evaluation) rather than relying solely on operational evaluation of model predictions under base line conditions. Here simulations from the Community Multiscale Air Quality (CMAQ) modeling system were conducted for 2002 and 2005, a time period characterized by significant emissions reductions associated with the EPA's Nitrogen Oxides State Implementation Plan Call (NO_x SIP Call) as well as mobile sources. These simulations were used to evaluate the performance of different forms of the RRF metric for projecting 2002 to 2005 against 2005 observed ozone levels. The evaluation study showed that the current form of the RRF calculation is generally well designed for predicting the future 8-hr ozone "design value" metric used for determining attainment. Specifically, the methodology of using air quality model simulations in a relative sense provided better estimates of future ozone design values than using the modeled future year simulation alone. Alternative forms of the RRF metric were found to be very similar to the current methodology in terms of evaluation metrics. However, alternative RRF metrics were sensitive to the number of days used in the calculation of the RRF. Approaches which used more days in the RRF calculation (relative to the 2007 guidance approach) had slightly higher bias and error in predicting 2005 design values compared to approaches using only a subset of high ozone days.

Keywords: Dynamic Evaluation; NOx SIP Call; ozone NAAQS; CMAQ version 5.0.1; relative response factor

1 Introduction

The U.S. Environmental Protection Agency sets National Ambient Air Quality Standards (NAAQS) for six criteria pollutants considered harmful to public health and the environment. When an area is found to exceed these standards and is subsequently designated as a "nonattainment area" for a particular pollutant, states are required by the Clean Air Act to develop State Implementation Plans (SIP) to describe what emission control strategies will be used to attain the NAAQS by a target date. For ozone, an important aspect of SIP development is the use of air quality models to demonstrate that the planned emission controls will provide the necessary

reduction in pollutant levels to achieve attainment of the standard. In 1999 the EPA provided draft guidance on the use of air quality models for SIP development for the 8-hr NAAQS ozone standard (EPA, 1999). The guidance outlined a method for using the ratio of model output from future-case and base-case emission conditions under fixed meteorological conditions to calculate a Relative Response Factors (RRF). The RRF is then used to scale observed ozone mixing ratios representative of the base year to determine if future year ozone levels will be in compliance with the NAAQS. This use of the model in a relative sense was a different methodology than the previous guidance for the 1-hr ozone standard which relied only on absolute modeled future-case concentrations. The RRF approach was updated and finalized in the 2007 guidance document (EPA, 2007).

One motivation for the use of air quality models in a relative approach is that anchoring the future year prediction on an observed base year value is expected to reduce problems posed by imperfect model performance (EPA, 2007). That is, some types of systematic model errors are expected to "cancel out" using the relative approach. In addition, several studies have shown that the relative approach is much less sensitive to the choice of modeling system or model configuration compared to using the model output directly (Jones et al., 2005; Sistla et al., 2004).

The use of photochemical grid models to predict changes in air quality motivates the need for additional evaluation, beyond the standard operational evaluation of model predictions under base line conditions. Hogrefe et al. (2008) showed that base model evaluation does not necessarily inform how well the model will estimate the change in air quality due to changes in emissions. Here a dynamic evaluation of the Community Multiscale Air Quality Modeling system (CMAQ) was conducted using a retrospective case study to evaluate the performance of different forms of the RRF metric to accurately predict reductions in ozone levels associated with reductions in NO_x emissions. The case study was based on modeling 2002 and 2005, a time period characterized by significant reductions in ozone precursor emissions. One contributor to these emission reductions was the EPA's Nitrogen Oxides State Implementation Plan Call (NO_x SIP Call) rule, fully implemented by 2004, and designed to reduce NO_x emissions from power plants in the eastern US (Gilliland et al., 2008). In addition, mobile source NOx emissions also decreased by roughly 20% during this time period due to motor vehicle fleet turnover to newer lower emitting vehicles (Foley et al., 2014). The combined emissions reductions contributed to a decrease in observed ozone levels up to 20-30% in the eastern US from 2002 to 2005. Additional background on the dynamic evaluation of CMAQv5.0.1 can be found in Part I of this study (Foley et al., 2014). The analysis in Part I shows that this modeling system tends to underestimate changes in ozone between these two years, similar to the finding in other studies (e.g. Gilliland et al., 2008; Zhou et al., 2013). The change in model estimated air quality was found to be most dependent on changes in emissions and somewhat dependent on changes in meteorology from 2002 to 2005. Here we expand on the dynamic evaluation by evaluating model projection metrics used to support regulatory modeled attainment demonstrations for the 8-hr Ozone NAAQS.

Section 2 provides details on the model simulations conducted for this case study including the development of a model simulation using 2005 emissions under 2002 meteorological conditions which was needed for the calculation of the RRF ratios. Section 3 presents a comparison of model-projected 2005 DVs using model output for base and future scenarios through the calculation of RRF metrics to an approach based solely on model predictions from the "future"

year simulation. While EPA guidance and other studies have recommended a relative approach for over a decade, to our knowledge, this is the first time that the difference between these two approaches has been quantified using a retrospective dynamic evaluation study. Five different forms of the RRF metric are compared in section 3 followed by a discussion in Section 4.

2 Model approach for NOx SIP Call time period

The model simulations used in this analysis are described in detail in sections 2 and 4 in Foley et al. (2014). In summary, four CMAQv5.0.1 simulations were performed over the continental U.S. using a grid with 12km horizontal resolution and 35 vertical layers. Meteorological inputs were based on WRF3.3 with MCIPv4.0 (see Appel et al., 2013 for further details). Emission inputs were developed based on 2002 and 2005 National Emission Inventory data using the Sparse Matrix Operator Kernel Emissions version3.1 emissions processing system (SMOKE; Houyoux et al., 2000), including year specific data for large point sources and year specific mobile emissions derived from MOVESv2010b. Boundary conditions were based on 2005 monthly median values from a GEOS-Chem v9-01-02 simulation (http://wiki.seas.harvard.edu/geos-chem/) using v8-02-01 chemistry, GEOS-5 meteorology and ICOADS shipping emissions (Henderson et al., 2014).

Two base year simulations were conducted for June 1 through September 30th 2002 and 2005 (Sim02e02m, Sim05e05m). In addition, two "cross" simulations were used to simulate air quality under 2005 emissions with 2002 meteorology (Sim05e02m) and 2002 emissions with 2005 meteorology (Sim02e05m). Modeling for attainment demonstrations is based on a base year and a future year where the meteorology is held constant across both simulations. Thus the focus of this analysis was on the Sim02e02m and Sim05e02m simulations. The other two simulations are used briefly to illustrate the change in ozone levels due to changes in meteorology from 2002 to 2005.

Many methods exist for creating future year emissions for attainment demonstrations and different methods have been developed for different emission sources. Here the NEI-based 2005 emission estimates are modified, as described below, to create the emission inputs for Sim05e02m which is used as the 'future' simulation in the RRF calculations. The focus of the current study is to isolate the errors and biases in predicted ozone levels that are associated with either the form of the RRF metric or the air quality modeling system itself. The amount of error in the predicted future case ozone levels that can be attributed to errors in projecting future emission levels will depend on the methodology used. An analysis of this type of modeling uncertainty is beyond the scope of the current work.

For the simulation with 2005 emissions and 2002 meteorology (Sim05e02m), the emissions from electrical generating units (EGUs) with available continuous emission monitoring systems (CEM) data in 2005 were adjusted to account for the impact of different meteorology in 2002. Summertime 2005 NO_x emissions are generally lower than 2002 emissions due to emission reductions. Temporal fluctuations are different due to differences in electricity demand which is heavily influenced by year-specific meteorology. 2005 emission levels with 2002 meteorological patterns (EMIS05e02m) were estimated by scaling the hourly 2002 CEM emissions (CEM2002) based on the ratio of summer total CEM emissions (STCEM) for a particular EGU unit in 2005 versus 2002: $EMIS05e02m = CEM2002 \times (STCEM2005/STCEM2002)$. An analogous calculation is made to estimate 2002 emissions with 2005 meteorological patterns for Sim02e05m. Emissions from electrical generating point sources with no CEMS data used 2005 annual total

emissions scaled with 2004-2006 annual-to-month ratios and 2002 day-to-month ratios in order to match the daily temporal fluctuations of the point sources with CEMS data in 2002. Additional details on the calculation of the scaling ratios is provided in Supplemental material S2.

Mobile emissions for Sim05e02m were based on a MOVES simulation using 2005 VMT and emission factors and 2002 meteorological inputs. Emissions from nonroad (e.g. construction), industrial point and commercial marine sectors are based on 2005 emission levels but shifted to match the day-of-the week in 2002. Emissions from fertilizer application, biogenic sources, NO_x from lightning, fires and dust are based on 2002 estimates since these sources are associated with meteorological conditions. All other sectors have identical emissions inputs for both Sim02e02m and Sim05e02m.

Section 3 outlines how the base year (Sim02e02m) and future year (Sim05e02m) simulations were used to calculate different RRF metrics in order to estimate 2005 ozone "design values". As part of determining which areas are attaining the ozone standard, EPA calculates ozone design values (DVs) as the 3-year average of the annual 4th highest maximum daily 8-hour average (MDA8) ozone concentration. Determining attainment based on multiple years of observations reduces the impact of interannual meteorological variability on pollutant concentrations. For attainment demonstrations in which states show how they plan to bring future ozone levels in line with the standard, EPA recommends establishing site-specific base DVs to be the average of three consecutive 3-year DVs, centered about the modeling year (which is effectively a 5-year weighted average DV) to further reduce the efforts of meteorological variability on design values. However, in this retrospective study the base and 'future' years are very close together in order to isolate the impact of a specific emission control event (i.e. the NO_x SIP Call). As a result, the standard calculation of the 2002 DVs would have been based on summers both before and after the controls were implemented in 2003 and 2004. To avoid this issue, the following analysis is based on 3-year DVs using 2000-2002 as the base year average (2002 DV) and 2004-2006 as the future year average (2005 DV).

Ozone observations from 644 monitoring stations from the EPA's Air Quality System (AQS; http://www.epa.gov/ttn/airs/airsaqs) across the U.S. were used to evaluate model predicted 2005 design values. These monitoring sites were selected from the entire AQS network based on observed 2002 DVs. Only sites with 2002 DVs greater than 75ppb were included in this study. Sites with lower base year ozone are of less interest because these areas would not be expected to have attainment issues for this case study.

3 Comparison of 5 RRF metrics

In this section the form of the 8-hr ozone NAAQS attainment test is reviewed followed by a comparison of alternate RRF metrics. The RRF metrics are also compared to using the model in an absolute sense rather than in a relative sense.

The 8-hr ozone model attainment test for monitoring site *i* provides an estimate of the future DV for the year in which attainment is required (DVF_i) based on the relative response factor (RRF_i) and the base design value for the site (DVB_i):

$$DVF_i = RRF_i \times DVB_i, \quad (1)$$

where RRF_i is a ratio of modeled future ozone to modeled base ozone. In the model attainment test suggested by the 2007 guidance (EPA, 2007), the modeled base ozone value (the denominator of the RRF) is an average of MDA8 ozone on all days in the base simulation greater than or equal to 85ppb. If there are less than ten days ≥85ppb in the base year, then the average of the ten highest days in the simulation are taken as long as each of these values is ≥70ppb. If there are at least five days but less than ten days with MDA8 ozone ≥70 ppb then the average of all days ≥70ppb are used. In the case when there are less than five days with MDA8 ozone ≥70ppb the RRF is not calculated. The modeled future ozone (the numerator of the RRF) is the average of the MDA8 ozone in the future year simulation over the same calendar days used in the base ozone calculation. This is referred to as the TH85 ("Threshold 85") approach for the remainder of the paper.

Before exploring alternate methods for calculating the RRF at a given site, the NO_x SIP Call case study was used to evaluate two underlying assumptions of the ozone attainment test. First is the idea that the change in the average of high ozone values is a good predictor of the change in the 4th highest ozone value across several summers. The motivation for using a threshold approach is that days with high base-case ozone levels are expected to be the days most representative of nonattainment conditions, which is what emission controls are designed to limit. Specifically, high base-case ozone levels have been found to be more responsive to emission reductions compared to days with lower ozone (Hogrefe et al., 2008). To investigate this with the current dataset, "daily RRF" values were calculated for each summer day in 2002 (June – September). The daily RRF at a given grid cell containing an AQS monitor is the ratio of the ozone from the 2005 simulation using 2002 meteorology (Sim05e02m) to the ozone value from the 2002 simulation (Sim02e02m) at that grid cell. That is, for location i, day j, $dailyRRF_{ij} = \frac{MDA8_{ij}^{05e02m}}{MDA8_{ij}^{02e02m}}$. Figure 1 shows daily RRF values as a function of 2002 modeled MDA8 ozone values. The top left figure shows that the daily RRF values across all sites tend to steadily decrease with increasing base level ozone values and then level off around 80ppb. The daily RRFs for several urban sites are also shown with the individual days that would be used in the calculation of the TH85 RRF shown in red.

To further evaluate the use of a threshold approach, an "observed" RRF was calculated at each AQS site based on the ratio of the average of observed MDA8 ozone in June-September 2005 to 2002. At site i, the denominator of the ratio is the average of the N_i days in 2002 with observed MDA8 ozone ≥85ppb and the numerator is the average of the top N_i MDA8 observed ozone days in 2005. The value of N_i is the same in the calculation of the numerator and denominator at a given site but can change across AQS sites. For example, if there were 15 observed MDA8 ozone values greater than 85ppb at a particular AQS site in 2002, then the denominator is the average of these 15 MDA8 ozone values and the numerator is the average of the top 15 observations at that site in 2005. An observation based estimate of the 2005 DV was then calculated by multiplying the RRF times the 2002 DV. Figure 2(a) shows the observed RRF approach does indeed evaluate well against the actual observed 2005 DVs. Note that unlike the model-based RRF, this calculation captures changes in both meteorology and emissions across these years. This result is included to provide context for the evaluation metrics used to evaluate the model-based results. Using a model-based RRF approach would not be expected to improve upon the observation-based metrics (e.g. MB = -1.9ppb, RMSE=5.6ppb, R^2 = 0.68).

The second issue evaluated was whether it is more appropriate to use the model output in a relative sense rather than comparing the absolute future year model prediction (e.g. output directly from Sim05e02m) to the level of the 8-hr ozone NAAQS. The second and third panels in Figure 2 show the observed 2005 DV compared to (b) the 4th highest MDA8 ozone value from Sim05e02m and (c) the predicted 2005 DV using the model-based TH85 RRF approach. Using the TH85 RRF approach results in smaller positive bias (3.8ppb vs. 4.1ppb) and much higher correlation (R^2 = 0.67 vs. 0.37) compared to the Sim05e02m prediction. The implication for attainment testing is that the relative approach anchored to 2002 observed design values is better able to predict exceedances of both the 75ppb (threshold for current 8-hr O_3 NAAQS) and 84ppb (threshold for NAAQS during 1997-2008) as reflected by the accuracy scores (Acc75 and Acc84). Note that the 1997 NAAQS is 0.08ppm with DVs rounded to two significant digits, i.e., any DV below 0.085ppm (85ppb) is considered in attainment. The 2008 standard is 0.075ppm with DVs truncated to three significant digits, i.e., any DV below 0.076ppm (76ppb) is considered in attainment. These results indicate that for the current O_3 NAAQS, the methodology of using high base-case ozone days and using the air quality model simulations in a relative sense is well suited for determining attainment and can mitigate some biases in the modeling system being applied.

Several alternate RRF calculations were compared in order to see if an alternate approach could be adopted to improve the accuracy of model-based DV predictions. To test whether it makes sense to tie the RRF calculation to the level of the ozone NAAQS, two additional threshold approaches were compared to the TH85 RRF. TH76 is based on averaging model days in 2002 with MDA8 ozone ≥76ppb. If there are fewer than ten days ≥76ppb in 2002 then all days ≥60ppb are used. If there are fewer than five days ≥60ppb then this metric is not calculated. TH71 is constructed in an analogous manner based on days greater than or equal to 71ppb. To avoid the use of a threshold all together, two "number of days" tests were also compared. The Top10 RRF averages the ten highest MDA8 ozone days in 2002 that are above 60ppb. If there are less than ten days ≥60ppb in 2002 then all days ≥60ppb are used. If there are less than five days ≥60ppb then this metric is not calculated at that site. The Top20 RRF is constructed in the same fashion except that at the first step the twenty highest MDA8 ozone days are used. The numerator of the RRF for all of these metrics is calculated by averaging the MDA8 ozone in the future year simulation over the same calendar days that were used in the base ozone calculation. Table 1 provides an overview of the five metrics that were compared.

Table 2 provides evaluation statistics for the five methods compared to using the output from the future model simulation directly (i.e., Sim05e02m). The 644 AQS sites are subset to sites where all six methods could be calculated. Twenty-five of these sites did not have five days with MDA8 ozone ≥70ppb needed for the calculation of the TH85 metric, reducing the final dataset to n=619. (Note that Tables S1 and S2 in the Supplemental material also show these statistics segregated by NOx SIP call states and non-NOx SIP Call states.)

The evaluation statistics of the five RRF type methods are almost identical with the TH85 and Top10 metrics having slightly lower bias and RMSE compared to the other 3 RRF metrics. All of the methods have a positive bias (3.8-4.2ppb) which means the model did not predict as large a decrease in the ozone DVs from 2002 to 2005 as was observed. Part of this bias can be attributed to the fact that the model based predictions do not account for changes in meteorology across these

summers while the observed changes in DV, though averaged over three years, do not completely eliminate the effects of such changes in meteorology. Analysis of the second cross run, Sim02e05m, showed that the modeled change in meteorology across these years did lead to a decrease in ozone levels in many parts of the domain (see Supplemental Figure S1). For example, using the Sim05e05m output for the numerator of the model-based RRF calculations (i.e. creating a model RRF that captured both meteorological and emissions changes) reduced the mean bias in the DV predictions to 2.4ppb as shown in Figure 2(d). The remaining underestimation of the observed change is a result of the model predictions for high summer time MDA8 ozone being too low in 2002 and too high in 2005 (the dynamic evaluation of these two years is discussed in detail in Part I, Foley et al. 2014). In other scenarios when the bias in the base and future year simulations are more comparable (e.g. the MDA8 ozone predictions are too high in both years) using the relative approach would be expected to provide an even greater improvement in prediction performance compared to using a single model simulation in an absolute sense.

Table 2 also provides two metrics for evaluating the 2005 predictions in a categorical sense. The modeled 2005 DVs were compared to the ozone NAAQs level to determine whether a site exceeded the standard or was in attainment. The model prediction at a given site was compared to the observed 2005 DV, producing four outcomes: true exceedance, false exceedance, false attainment, true attainment. For example, for the 84ppb NAAQS if *Mod* is the model predicted 2005 DV and *Obs* is the observed 2005 DV, the four outcomes are: *Mod* ≥85ppb and *Obs* ≥85ppb (true exceedance); *Mod* ≥85ppb and *Obs* <85ppb (false exceedance); *Mod* <85ppb and *Obs* ≥85ppb (false attainment); *Mod* <85ppb and *Obs* <85ppb (true attainment). Spatial plots of these categorical outcomes for the future model output and the TH85 RRF approach are shown in Figure 3 and Supplemental Figure S3 for both the 84ppb and 75ppb standards for the entire modeling domain and the Eastern U.S., respectively. Note that the locations where false attainments or false exceedances occurred are specific to model simulations used in this case study and would be expected to change for a different model setup or a different set of simulation years.

The accuracy of the model predictions is the fraction of sites where the model produced a correct prediction (true exceedance or true attainment). For the 75ppb standard, the accuracy of all of the RRF methods is higher than the future model output (0.78-0.79. vs. 0.72). The accuracy scores for the 84ppb standard are even higher with the TH85 and Top10 metrics having the highest score (0.82) of all of the methods. Figure 3 shows the TH85 approach decreases both false attainment predictions (17 vs. 28) and false exceedance predictions (97 vs 114) compared to the future model. Because the TH85 RRF approach defaults to the top ten days when there are ten or fewer modeled days ≥85ppb, the Top10 RRF approach and the TH85 approach were identical at 77% of the AQS sites. At the remaining sites, the Top10 RRF had slightly lower bias (MB of 5.3ppb vs. 5.4ppb) and error (RMSE of 7.9ppb vs. 8.0ppb) than the TH85 approach. Moreover, the Top10 RRF and the TH85 RRF approaches predict the same attainment designations (i.e. exceedance or attainment) at all but a single AQS site for both the 84ppb and the 75ppb standards.

This analysis showed that all of the tested RRF metrics were similar. However methods that used a larger of number of days in the RRF calculation (see Figure 4) all led to higher MB and RMSE compared to the Top10 RRF and TH85 RRF approaches. The Top10 metric was found to be a comparable metric to the current TH85 approach and offers the advantage that it is independent of the threshold used in the NAAQS standard, which could be revised again in the future. In

addition, the Top10 metric simplifies the RRF calculation and ensures that only the top percentile of ozone days are included in the model averages. For example, some locations with very high ozone had more than twenty or even thirty days above 84ppb in 2002, i.e. 15-25% of the entire simulation period.

4 Summary

The dynamic evaluation case study based on the 2002-2005 time period characterized by emissions reductions associated with the NOx SIP Call and mobile source emissions offered a valuable opportunity to assess the methodology used in the ozone attainment test, as well as the suitability of the CMAQ modeling system for use in this type of regulatory modeling application. The RRF approach was found to be generally well designed to capture changes in the observed DVs across these years, and offered several advantages over using output from the air quality model directly, in terms of lower bias, higher correlation and higher accuracy in predicting ozone exceedances. There were no large differences found in the evaluation of the five different RRF metrics. Basing the metrics on the highest modeled base-year ozone days yielded a slightly lower bias and higher correlation compared to the metrics that averaged over a larger percentage of the simulation days.

The performance of the RRF based approach was quantified here based on aggregate statistics for monitoring sites across the contiguous US domain. For a given SIP there may be local circumstances where large uncertainties in model output and observations would necessitate analysis beyond the RRF methodology, referred to by the EPA guidance as a "weight of evidence" analysis. For example, Vizuete et al. 2010 and 2011 provide detailed discussion and evaluation of RRF calculations for the Houston, TX area. Such area-specific analysis will continue to be an important part of the SIP process regardless of the specific RRF methodology being applied.

This study also demonstrated how inter-annual meteorological variability can impact the attainment test and be a confounding factor in the dynamic evaluation of the air quality modeling system being used. In this case, the change in meteorological conditions from 2002 to 2005 caused ozone to decrease in many locations resulting in an increase in the positive bias in the model predicted DVs. The use of 3-year DVs helps to reduce the impacts of year-to-year changes in meteorology on observed pollutant levels but likely does not eliminate these effects. Future work using longer-term records of observations and model simulations is needed to assess whether using model output from multiple years in calculating RRFs may improve the ability of the modeling system to reproduce observed changes in DVs.

Disclaimer: Although this work was reviewed by EPA and approved for publication, it may not necessarily reflect official Agency policy.

Acknowledgements

The authors would like to recognize the many contributions of others in their support in processing the large suite of emission, meteorology and boundary condition inputs needed for the CMAQv5.0.1 simulations, in assisting with observed datasets and in providing many thoughtful suggestions on how to improve the final analysis: Wyat Appel, Rob Gilliam, Jim Godowitch, Rob Pinder, George Pouliot, Shawn Roselle (EPA/ORD/NERL); Kirk Baker, Alison Eyth, Sharon Philips, Benjamin Wells, Alexis Zubrow (EPA/OAR/OAQPS); Allan Beidler, Lucille Bender,

Ryan Cleary (Computer Sciences Corporation); Barron Henderson (now at Univ. of FL); and Farhan Akhtar (now at U.S. State Dept.).

References

Appel, K.W., Pouliot, G.A., Simon, H., Sarwar, G., Pye, H.O.T., Napelenok, S.L., Akhtar, F., Roselle, S.J. (2013) Evaluation of dust and trace metal estimates from the Community Multiscale Air Quality (CMAQ) model version 5.0. *Geoscientific Model Development* 6, 883-899.

Foley, K.M., Hogrefe, C., Pouliot, G., Possiel, N., Roselle, S., Simon, H., Timin, B. (2014) Dynamic evaluation of CMAQv5.0.1 Part 1: Quantifying changes in ozone associated with the EPA's NOx SIP Call Rule, *in review.*

Gilliland, A.B., Hogrefe, C., Pinder, R.W., Godowitch, J.M., Foley, K.M., Rao, S.T. (2008) Dynamic evaluation of regional air quality models: Assessing changes in O3 stemming from changes in emissions and meteorology. *Atmospheric Environment* 42, 5110-5123.

Henderson, B.H., Akhtar, F., Pye, H.O.T., Napelenok, S.L., Hutzell, W.T. (2014) A database and tool for boundary conditions for regional air quality modeling: description and evaluation. *Geoscientific Model Development*, 7, 339-360.

Hogrefe, C.; Civerolo, K.L.; Hao, W., Ku, J.Y.; Zalewsky, E.E.; Sistla, G. (2008) Rethinking the assessment of photochemical modeling systems in air quality planning applications. *J. Air & Waste Manage. Assoc.*, 58, 1086-1099.

Houyoux, M. R., Vukovich, J. M., Coats Jr., C. J., Wheeler, N. J. M., Kasibhatla, P. (2000) Emission inventory development and processing for the seasonal model for regional air quality, *Journal of Geophysical Research: Atmospheres*, 105, 9079–9090.

Jones, J.M.; Hogrefe, C.; Henry, R.F.; Ku, J.-Y.; Sistla, G. (2005) An assessment of the sensitivity and reliability of the relative reduction factor (RRF) approach in the development of 8-hr Ozone Attainment Plans; *J. Air & Waste Manage. Assoc.*, 55, 13-19.

Sistla, G.; Hogrefe, C.; Hao, W.; Ku, J.-Y; Zalewesky, E.; Henry, R.F.; Civerolo, K. (2004) An operational assessment of the application of the relative reduction factors (RRF) in demonstration of attainment of the 8-hr ozone National Ambient Air Quality Standard (NAAQS); *J. Air & Waste Manage. Assoc.*, 54, 950-959.

United States Environmental Protection Agency, 1999. Draft report on the use of models and other analyses in attainment demonstrations for the 8-hour ozone NAAQS. EPA-44/R-99-0001, May 1999, *United States Environmental Protection Agency*, Research Triangle Park, NC 27711.

United States Environmental Protection Agency, 2007. Guidance on the use of models and other analyses for demonstrating attainment of air quality goals for ozone, PM2.5, and regional haze. EPA-454/B-07-002, April 2007, *U.S. Environmental Protection Agency*, Research Triangle Park, NC, 27711.

Vizuete, W., Biton, L., Jeffries, H., Couzo, E. (2010) Evaluation of relative response factor methodology for demonstrating attainment of ozone in Houston, Texax. *J. Air & Waste Manage. Assoc.*, 60, 838-848.

Vizuete, W., Jeffries, H.E., Tesche, T.W., Olague, E.P., Couzo, E. (2011) Issues with ozone attainment methodology in Houston, TX. *J. Air & Waste Manage. Assoc.*, 61, 238-253.

Zhou, W., Cohan, D.S., Napelenok, S.L. (2013) Reconciling NOx emissions reductions and ozone trends in the U.S., 2002-2006, *Atmospheric Environment*, 70, 236-244.

List of Figures

Figure 1. Daily RRF ratios versus 2002 modeled MDA8 ozone values at all AQS sites (top left) and at 5 select AQS sites in very urban areas. The color scale of the density scatter plot in the top right figure represents the percent of the data that fall within a particular pixel in the plot, e.g., the red pixels in the plot indicate areas where more than 1.5% of the data (>10 data points) would be clustered in a regular scatterplot. The green curve shows a spline fit to the daily RRF values. The red points in the scatter plot indicate what days were used in the calculation of the TH85 RRF metric for each site.

Figure 2. Comparison of 2005 observed versus predicted DV at n=619 AQS sites. Predicted DVs are based on using the (a) observed TH85 RRF approach, (b) the 4th highest MDA8 ozone value from the Sim05e02m simulation (c) the modeled TH85 approach and (d) the modeled TH85 approach modified by using 2005 meteorology in the "future" model simulation (i.e. output from Sim05e05m rather than Sim05e02m is used in the numerator). Evaluation statistics include mean bias (MB), root mean square error (RMSE), R2, and accuracy scores for predicting attainment of the 75ppb and 84ppb NAAQS (Acc75, Acc84).

Figure 3. Observed 2002 and 2005 design values (top row) across the entire model domain. The remaining plots show model predicted 2005 DVs in a categorical sense based on output from Sim05e02m (left column) and the TH85 RRF based approach (right column) for both the 84ppb standard (middle row) and the 75ppb standard (bottom row). Four categories are depicted: true exceedance (green circle); false exceedance (purple triangle); false attainment (red inverted triangle); true attainment (yellow square).

Figure 4. The number of days used in the calculation of each of the five RRF metrics across the 619 AQS sites used in Table 2.

List of Tables

Table 1. Summary of five RRF metrics compared in section 3. TH85 was the method described in the 2007 guidance document.

RRF metric	Type of metric	Days averaged	Minimum ozone level used in the average	Minimum # days to be averaged
TH85	Threshold	Days ≥85	70	5
TH76	Threshold	Days ≥76	60	5
TH71	Threshold	Days ≥71	60	5
Top10	Number of days	Highest 10 days	60	5
Top20	Number of days	Highest 20 days	60	5

Table 2. Evaluation of different approaches for predicting the observed 2005 DVs at 619 AQS sites. The evaluation statistics include mean bias (MB), root mean square error (RMSE), R2, and accuracy for predicting attainment for the 75ppb and 84ppb NAAQS.

	MB (ppb)	RMSE (ppb)	R^2	Accuracy for 75ppb	Accuracy for 84ppb
Future Model	4.1	8.5	.37	.72	.77
TH85 RRF	3.8	6.6	.67	.79	.82
TH76 RRF	4.2	6.9	.66	.78	.80
TH71 RRF	4.2	6.9	.67	.78	.80
Top10 RRF	3.8	6.6	.67	.78	.82
Top20 RRF	4.1	6.8	.66	.78	.81

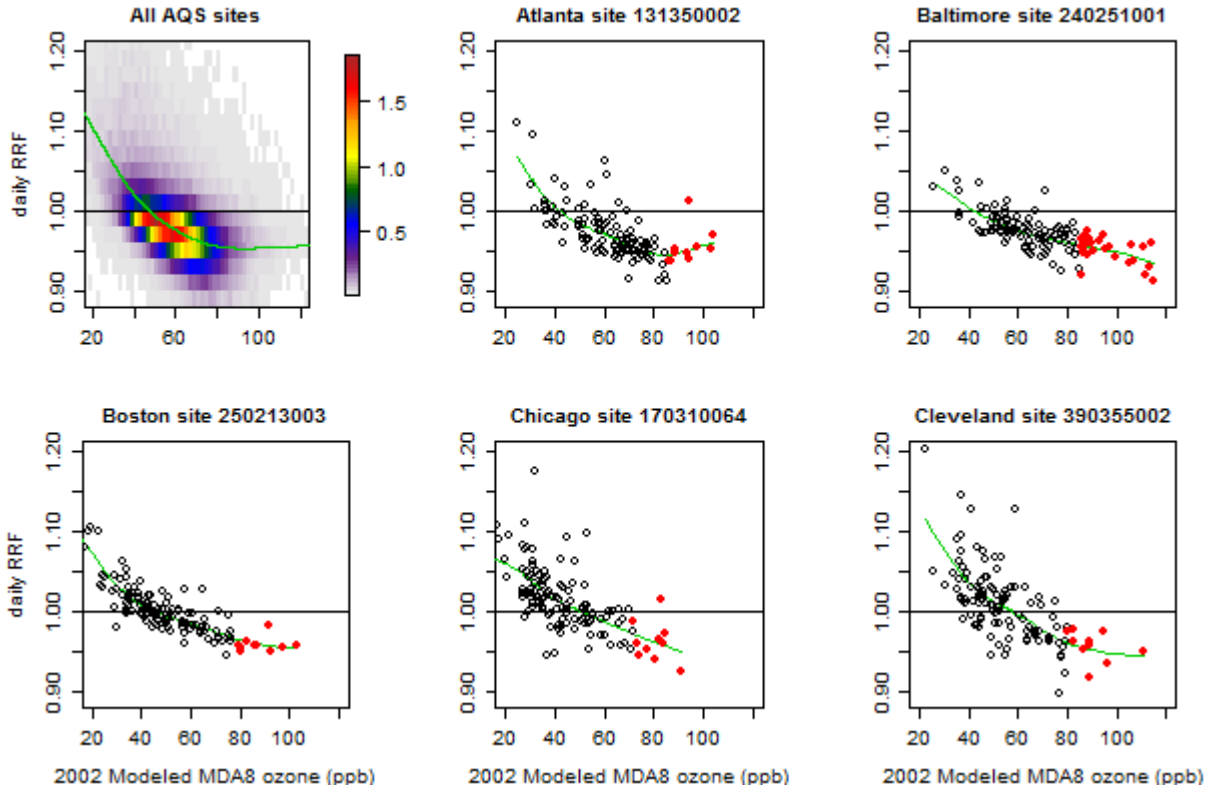


Figure 1. Daily RRF ratios versus 2002 modeled MDA8 ozone values at all AQS sites (top left) and at 5 select AQS sites in very urban areas. The color scale of the density scatter plot in the top right figure represents the percent of the data that fall within a particular pixel in the plot, e.g., the red pixels in the plot indicate areas where more than 1.5% of the data (>10 data points) would be clustered in a regular scatterplot. The green curve shows a spline fit to the daily RRF values. The red points in the scatter plot indicate what days were used in the calculation of the TH85 RRF metric for each site.

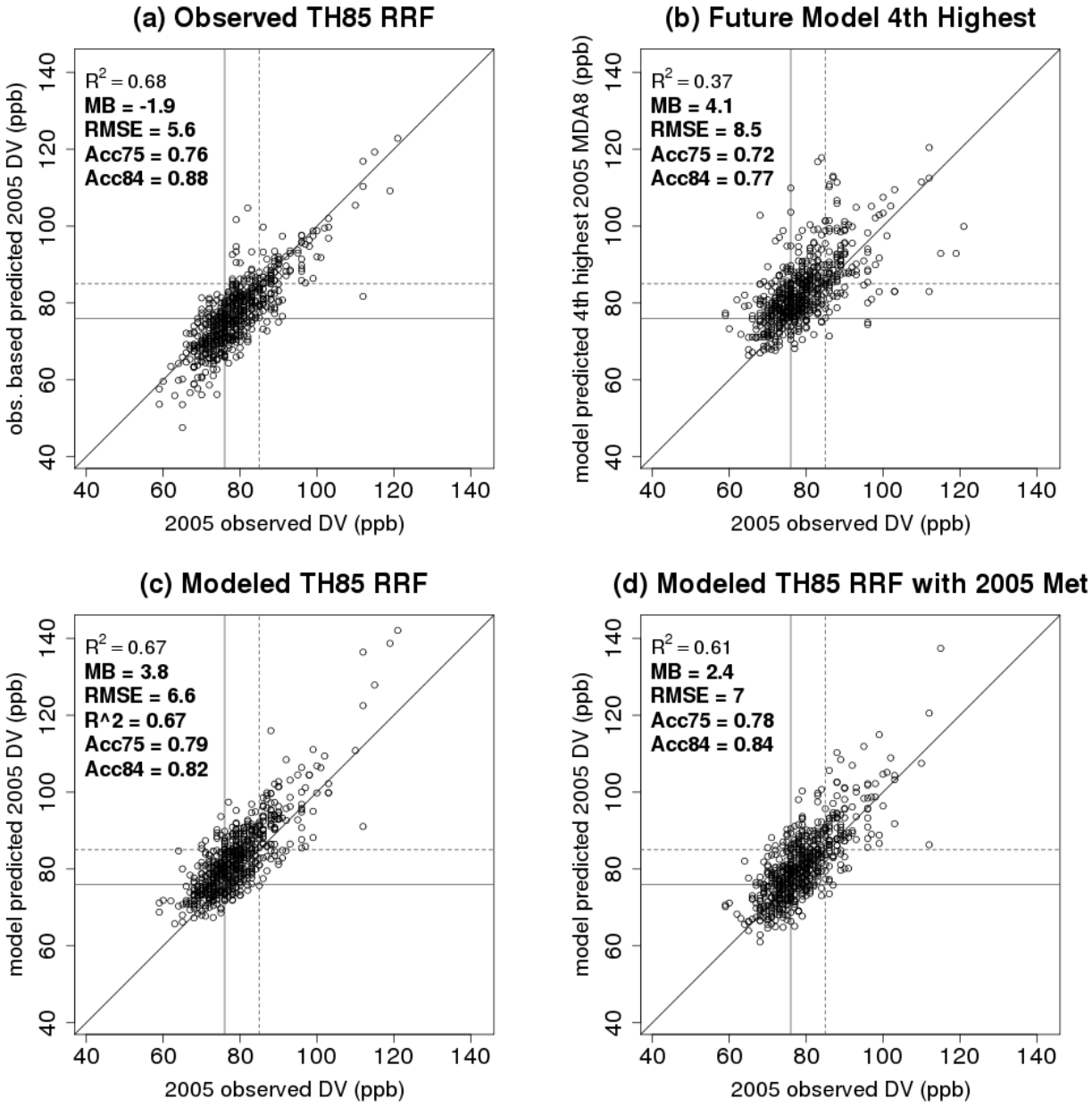


Figure 2. Comparison of 2005 observed versus predicted DV at n=619 AQS sites. Predicted DVs are based on using the (a) observed TH85 RRF approach, (b) the 4th highest MDA8 ozone value from the Sim05e02m simulation (c) the modeled TH85 approach and (d) the modeled TH85 approach modified by using 2005 meteorology in the "future" model simulation (i.e. output from Sim05e05m rather than Sim05e02m is used in the numerator). Evaluation statistics include mean bias (MB), root mean square error (RMSE), R2, and accuracy scores for predicting attainment of the 75ppb and 84ppb NAAQS (Acc75, Acc84).

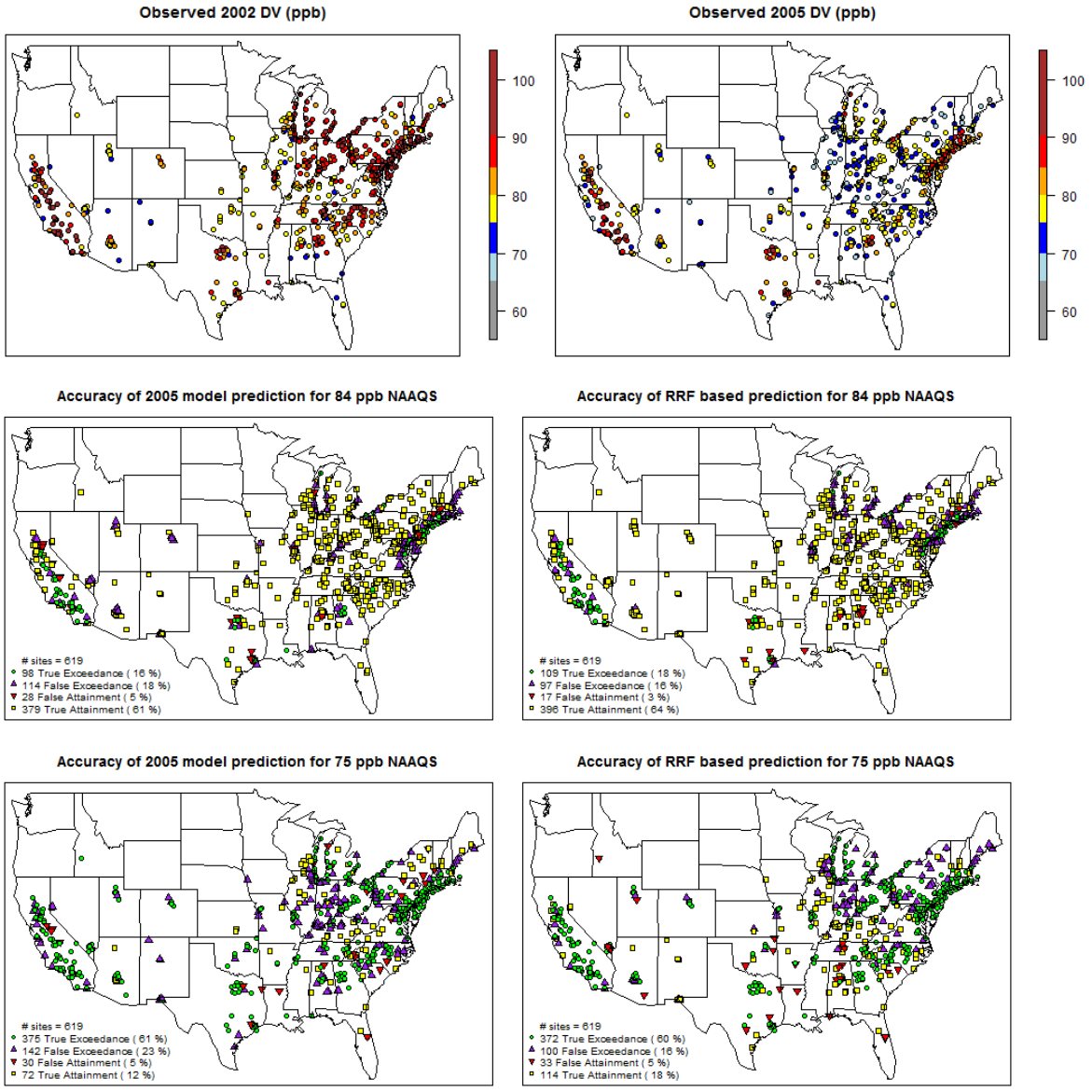


Figure 3. Observed 2002 and 2005 design values (top row) across the entire model domain. The remaining plots show model predicted 2005 DVs in a categorical sense based on output from Sim05e02m (left column) and the TH85 RRF based approach (right column) for both the 84ppb standard (middle row) and the 75ppb standard (bottom row). Four categories are depicted: true exceedance (green circle); false exceedance (purple triangle); false attainment (red inverted triangle); true attainment (yellow square).

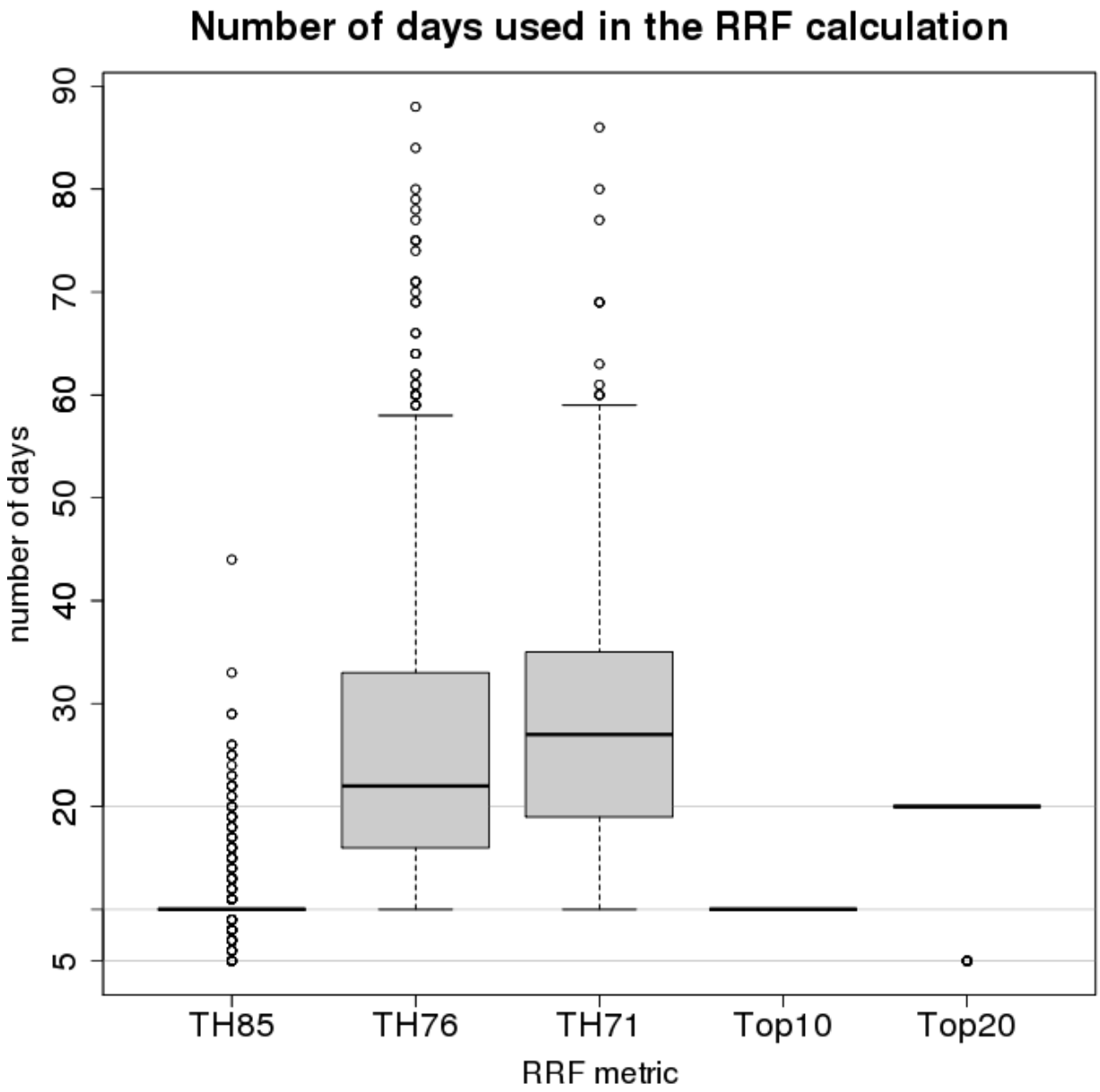
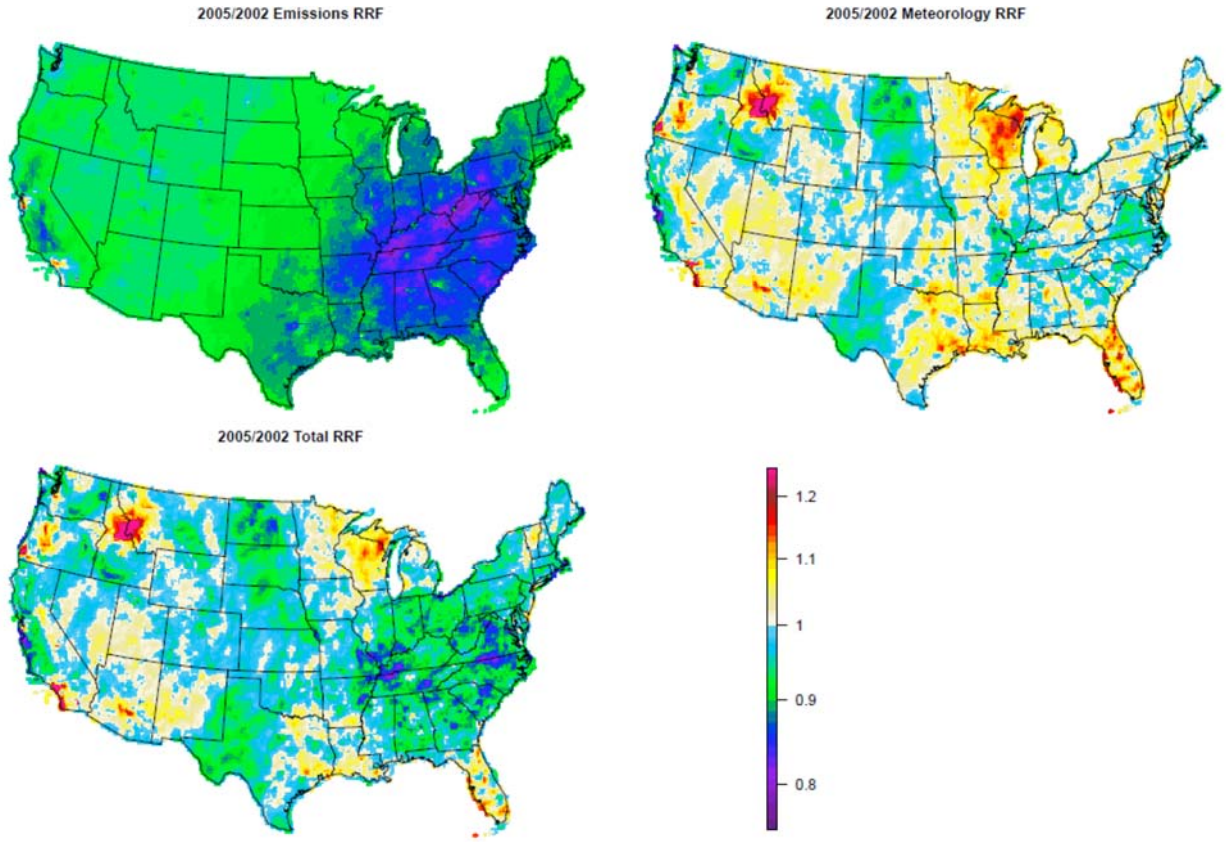


Figure 4. The number of days used in the calculation of each of the five RRF metrics across the 619 AQS sites used in Table 2.

Supplemental Material

Figure S1. Top10 RRF values for the entire US domain. The top left figure is based on the ratio of Sim05e02m output to Sim02e02m output, representing the RRFs used in the attainment evaluation in this paper. The top right figure shows the ratio of Sim02e05m output to Sim02e02m. This figure represents the model-predicted change in ozone due to changes in meteorology across these two years under 2002 emission levels. The final figure on the bottom row is a "total" RRF based on the ratio of Sim05e05m to Sim02e02m output and shows the model predicted change in ozone levels from 2002 to 2005 due to changes in both emissions and meteorology.



S2. Creation of hourly emissions for point sources without CEMS data (ptipm)

Pseudo-CEMS data were created for point sources that only have annual total emissions available in the NEI. For Sim02e02m, state-specific month-to-annual ratios were created by calculating three-year averages of 2001-2003 CEMS data for each month and dividing these monthly averages by the three year annual average for the state. The annual total ptipm emissions for each unit were then allocated to month totals using these state-specific monthly ratios. To allocate the monthly emissions to each day, state-specific day-to-month ratios were calculated using daily 2005 CEMS data divided by the monthly average for 2005. These state-specific day-to-month factors were then multiplied by the monthly total emissions for a given unit to calculate the total emissions on each day for that unit. The resulting daily emissions were input into the SMOKE processing system and hourly-to-daily allocation was performed using diurnal profiles. An analogous calculation was made to estimate hourly 2005 emissions for the Sim05e05m simulation based on 2004-2006 CEMS data. This is the standard method used in regulatory applications for creating simulations based on future emissions levels under current, base line meteorological conditions. Emissions inputs for Sim02e05m used 2002 ptipm unit annual total emissions scaled with 2001-2003 annual-to-month ratios to preserve the NO_x SIP call seasonal distribution and 2005 day-to-month ratios to preserve the meteorological patterns of the meteorological year. Emissions inputs for Sim05e02m used 2005 ptipm unit annual total emissions scaled with 2004-2006 annual-to-month ratios and 2002 day-to-month ratios.

Table S1. Evaluation of different approaches for predicting the observed 2005 DVs at 388 AQS sites within NOx SIP Call States. The evaluation statistics include mean bias (MB), root mean square error (RMSE), R^2, and accuracy for predicting attainment for the 75ppb and 84ppb NAAQS.

	MB (ppb)	RMSE (ppb)	R^2	Accuracy for 75ppb	Accuracy for 84ppb
Future Model	4.8	8.1	.35	.70	.80
TH85 RRF	4.6	6.4	.60	.78	.79
TH76 RRF	5.0	6.8	.59	.77	.77
TH71 RRF	5.1	6.8	.61	.77	.78
Top10 RRF	4.6	6.4	.60	.78	.79
Top20 RRF	4.9	6.6	.60	.77	.78

Table S2. Evaluation of different approaches for predicting the observed 2005 DVs at 231 AQS sites that are not within NOx SIP Call States. The evaluation statistics include mean bias (MB), root mean square error (RMSE), R^2, and accuracy for predicting attainment for the 75ppb and 84ppb NAAQS.

	MB (ppb)	RMSE (ppb)	R^2	Accuracy for 75ppb	Accuracy for 84ppb
Future Model	3.0	9.3	.37	.77	.71
TH85 RRF	2.5	7.0	.71	.79	.86
TH76 RRF	2.7	7.1	.71	.79	.85
TH71 RRF	2.7	7.1	.71	.79	.85
Top10 RRF	2.5	6.9	.71	.79	.86
Top20 RRF	2.6	7.0	.71	.79	.85

Figure S3. Observed 2002 and 2005 design values (top row) in the eastern US. The remaining plots show model predicted 2005 DVs in a categorical sense based on output from Sim05e02m (left column) and the TH85 RRF based approach (right column) for both the 84ppb standard (middle row) and the 75ppb standard (bottom row). Four categories are depicted: true exceedance (green circle); false exceedance (purple triangle); false attainment (red inverted triangle); true attainment (yellow square).