Available online at www.sciencedirect.com
ScienceDirect
Atmospheric Environment 42 (2008) 5403–5412
ATMOSPHERIC ENVIRONMENT
www.elsevier.com/locate/atmosenv

A procedure for inter-comparing the skill of regional-scale air quality model simulations of daily maximum 8-h ozone concentrations

John S. Irwin[a,*], Kevin Civerolo[b], Christian Hogrefe[c], Wyat Appel[d], Kristen Foley[d], Jenise Swall[d]

[a]*John S. Irwin and Associates, 1900 Pony Run Road, Raleigh, NC 27615, USA*
[b]*Division of Air Resources, New York State Department of Environmental Conservation, Albany, NY 12233, USA*
[c]*Atmospheric Sciences Research Center, State University of New York at Albany, Albany, NY 12222, USA*
[d]*Atmospheric Sciences Modeling Division[1], NOAA, Research Triangle Park, NC 27711, USA*

Received 13 January 2008; received in revised form 22 February 2008; accepted 23 February 2008

Abstract

An operational model evaluation procedure is described to quantitatively assess the relative skill among several regional-scale air quality models simulating various percentiles of the cumulative frequency distribution of observed daily maximum 8-h ozone concentrations. Bootstrap sampling is used to characterize the variability in the observed percentile values, thereby providing a means for assessing whether the differences seen between model predictions are significant. The procedure was designed to facilitate model inter-comparisons, since all that is needed to implement the procedure is for each modeler to provide a listing of the daily maximum 8-h ozone concentration predictions for a summer season for grid cells containing ozone monitors. Available ozone modeling results for the summer of 2002 from four regional-scale air quality simulations are used here to illustrate the results that can be obtained. These simulations were conducted using the Community Multi-Scale Air Quality (CMAQ) model with somewhat different setups. The modeling domains were different, but there is a region in the central Eastern United States where ozone estimates from all four simulations are available. Our objective is to describe the inter-comparison procedure, to illustrate the results obtained, and to stimulate discussions on how similar procedures might be developed and improved in the future.
Published by Elsevier Ltd.

Keywords: Model evaluation; Regional-scale; Air quality

1. Introduction

Envisioning atmospheric observations as resulting from partly deterministic and partly stochastic processes results from the fact that deterministic cause-and-effect is not sufficient to explain the observations (e.g., Lorenz, 1963, 1983; Julian and Murphy, 1972; Fox, 1984; Yoden, 2007). Atmospheric states exist for which very small changes in forcing can lead to different subsequent states. With this thought in mind, we envision the time-series of air quality measurements at any given location for

*Corresponding author. Tel.: +1 919 872 7654.
E-mail address: jsirwinetal@nc.rr.com (J.S. Irwin).
[1]In partnership with the US Environmental Protection Agency.

1352-2310/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.atmosenv.2008.02.046

a summer season to represent an individual realization from a population of possible outcomes. If we could experience this exact same summer season many times, the observed ozone values would "scatter" about an average value. Current models attempt to simulate what is to be seen "on average", but uncertainties arise due to limitations in our understanding of the relevant atmospheric processes and imperfect input data (e.g., meteorology, emissions, terrain, buildings, land-use). Thus, the deviations between observations and model predictions reflect naturally occurring variations (some of which may be stochastic) that current deterministic models do not explicitly simulate, as well as "uncertainties" arising from limitations in our knowledge and imperfect model input data.

Over the last several decades, there have been a series of workshops and position statements discussing the evaluation of air quality models and the importance of better characterization of their uncertainties, e.g., Hanna and Gifford (1971), Fox (1981, 1984), Demerjian (1985), Dabberdt et al. (2004), but aside from Dabberdt et al. (2004), these discussions have been more relevant to local-scale than to regional-scale air quality modeling systems. Suggestions for model evaluation methods have been provided to account for the fact that models do not predict stochastic variations seen in observations, but these also were developed primarily for local-scale models, e.g., Venkatram (1979, 1983), Weil et al. (1992), ASTM D 6589 (2000),[2] Dabberdt et al. (2004).

While the ASTM Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance (D 6589) is focused on local-scale models, the general principles discussed are relevant for any operational model evaluation method, namely: (1) model performance (or "skill") can be assessed through model inter-comparisons; (2) air quality models predict what is to be seen on average; thus, comparisons should be conducted with a spatial or temporal average of some feature in the observed concentrations; and (3) operational model evaluation procedures should provide a quantitative test of whether differences seen between models are significant.

Our purpose is to describe an operational model evaluation procedure that meets the above objectives and can be used to quantitatively assess the relative skill of different regional-scale air quality models in simulating the observed average daily maximum 8-h ozone concentration. Hopefully, this will stimulate discussions on how similar procedures might be developed and improved. Operational inter-comparisons provide increased communication among modeling groups, assist in timely identification and correction of model errors, encourage development of standardized benchmark calculations and promote the development of systematic records of progress in air quality modeling. Operational model evaluation methods involve running the full modeling system and comparing predicted and observed "features" in the meteorological and concentration spatial and temporal patterns. Such comparisons can reveal deficiencies, but diagnostic evaluation methods are needed to explore underlying causes for the differences seen and to assess where improvements are warranted in the modeling system.

In the following discussion, Section 2 outlines the model comparison procedure and the rationale for the manner in which comparisons are made of modeling results with observations and between modeling results. Section 3 discusses the data used to illustrate the model comparison procedure. Section 4 describes the analyses made to finalize the model comparison procedure and illustrates the comparison results that can be obtained. Section 5 provides a summary of the discussion.

2. Approach for operational modeling evaluation

2.1. What will we ask models to characterize?

Previous investigations (Hogrefe et al., 2001a, b, 2006; Biswas et al., 2001) have shown that current regional-scale air quality models have skill in replicating synoptic time-scale variations in the observations, but lack skill in replicating short-term variations (time scale of <11 h) in the observations. As discussed in Section 1, the lack of skill in replicating exactly what is observed is due to unresolved variations, which are envisioned to be partly deterministic and partly stochastic. Another reason for the differences seen between observations at a particular site and simulation results is the fact that the observations represent what is seen at a particular point, whereas the models predict volume average concentrations. Therefore, we expect models to properly characterize the large-scale variations, but we do not expect models to properly

[2]ASTM Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance (D 6589).

characterize the fine-scale and short-lived features. A goal of our model comparison procedure is to incorporate into the assessment procedure our "uncertainty" in characterizing from the available observations that which the models are attempting to simulate, namely what is to be seen "on average".

The United States Environmental Protection Agency (US EPA) design value[3] for 8-h average ozone is equal to the average of the 4th highest annual daily maximum 8-h average value for the most recent consecutive 3 years. A monitoring site must have a design value <85 parts per billion (ppb) to be considered to be in compliance with the ozone National Ambient Air Quality Standard. This ozone standard places emphasis on assessing regional-scale model's capability in simulating the extreme values in the daily 8-h ozone cumulative frequency distribution. The 4th highest value would be the 96th percentile value in the cumulative frequency distribution for a 92-day summer season.

To be in balance with the state-of-science in regional-scale air quality modeling systems, we examine the models' ability to replicate average percentile values (ranging from the 5th to the 95th percentiles) in the cumulative frequency distribution of daily maximum 8-h ozone values at each site for the summer season. We limit our comparisons to the 5th through the 95th percentiles, since we have only one summer season for analysis, which provides few values for characterization of the lowest and highest percentile values. For some purposes, the focus will be on the 95th percentile, but looking at other percentiles provides a more complete view of model performance and may reveal other strengths and weaknesses in a model run's performance. The 95th percentile is consistent with EPA's focus on the 96th percentile for NAAQS compliance purposes, as explained above.

The comparison procedure is to be conducted at each monitoring site within the domain that is common to all the modeling results being compared. This provides an assessment of relative modeling skill at each monitor, which can then be summarized over all monitors in the assessment.

[3]In the United States, a design value is the mathematically determined pollutant concentration at a particular site that must be reduced to, or maintained at or below the National Ambient Air Quality Standard to assume attainment. The design value tells us how a particular site or area compares with the National Ambient Air Quality Standards (NAAQS).

2.2. Motivation for bootstrap methods

A direct comparison of the selected percentile values could be done, but a more informed comparison would address our uncertainty in characterizing the average cumulative frequency percentile values from the available observations. This allows us to assess whether the differences seen between models are significant given the uncertainty we have in estimating the observed average percentile values. One means to estimate the uncertainty in population characteristics derived from an available set of observations is to employ bootstrap sampling.

Bootstrap sampling involves building pseudo-sets of observations by randomly sampling the available set of observations with replacement. An improvement over sampling with replacement from the entire set of 92 ozone 8-h average values observed over a summer season at each site is to sample 30 values with replacement from each month. This stratification captures month-to-month seasonal variations. When we select an observed day's 8-h average ozone value to be included in a bootstrap sample, we also select the corresponding modeling results for the selected day; this is called concurrent sampling in ASTM D 6589. Concurrent sampling links the model's simulation results to observations selected for inclusion in a given bootstrap sample. Concurrent sampling allows us to assess model skill and bias which may vary (e.g., as function of synoptic situation or average ozone concentration over the domain).

By randomly sampling each month's results with replacement, we are simulating the effects of experiencing a different collection of synoptic events within each month, and thus developing an estimate of the variability that might occur if we could experience this exact same summer season many times. The conclusions of the analysis are limited to the length of the data record and coverage of the ozone monitors used in the comparisons.

A question left for empirical resolution in Section 4 is how many bootstrap replications are needed in order for the final comparison results to be relatively insensitive to additional bootstrap replications being added to the analysis.

2.3. Block sampling

Autocorrelations in a time series of values adversely affects the estimates of means, cumulative frequency percentile values, and other similar

measures. In the context of a model evaluation exercise, understanding the variability of these estimates requires taking this correlation into account in our bootstrapping methodology. For instance, a larger variance associated with the estimates of population characteristics would imply that differences between modeled and observed characteristics would need to be larger before such differences would be deemed of consequence.

One means to capture the effects of day-to-day (serial) correlation in the observations when sampling a time-series of observations is to select short blocks (e.g., 2 consecutive days (pairs), 3 consecutive days (triplets) of values, see Wilks (1997) and ASTM D 6589 (2000)), rather than selecting independent samples with replacement. A question left for empirical resolution in Section 4 is the length of the block sample.

2.4. Comparing results from several model runs

On the basis of the following discussion, assume we have created 500 "summer seasons" of maximum daily 8-h ozone values, with 90 "days" in each summer season, having sampled 30 days with replacement from each of the summer months of June, July and August. Sorting each season's values from lowest to highest, we have 500 cumulative frequency distributions of 8-h ozone values (observations and estimates from each model run in the comparison).

Determining which of several model run predictions is best at characterizing an observed feature, and then whether results from other models differ significantly, is not trivial, as each model application has different confounding complications. Regional-scale photochemical models track and predict a large number of chemical species and a desirable definition of "best" would likely be that model run that routinely is closest to several chemical species deemed of importance simultaneously in comparison to other model runs. For this demonstration, we will make the simplest of comparisons at each monitoring site, namely, using only the percentile values of the cumulative frequency of maximum daily 8-h average ozone values. Admittedly, this offers the best chance of good correspondence resulting from compensating errors. However, our purpose is to illustrate an operational model evaluation technique for regional-scale models that is more quantitative than thus far has been discussed in the literature and that can be further developed in the future.

In this initial stage of seeing which of the modeling results is closest "on average" to what is observed, we will compute from each of the bootstrap "seasons" the absolute value of the difference between the modeled and observed percentile value, call this abs(Diff-1). We will define the best performing model run ("Base") for a particular percentile value, as the model run with the smallest median value for the 500 values of abs(Diff-1), where Diff-1 equals the observed percentile value minus the modeled percentile value. We chose to use the median as it is a robust estimate of the central tendency of a distribution of values and is less influenced than a mean value by outlying values in the tails of the distribution.

It is possible that even though the Base model has the smallest median for abs(Diff-1), it may not compare well with the observed percentile value. To investigate this, for each percentile of interest, we inspect the distribution of Diff-1 differences for the Base model. If the value of zero is encompassed between the 5th and the 95th percentile of the distribution of 500 Diff-1 differences, than we conclude that for the Base model run the difference in modeled and observed values for this percentile is of no consequence.

We know from the analysis of the 500 Diff-1 values that the Base model run is closest to what is observed "on average". What we wish to assess now is whether the results from another model run are essentially the same as those produced by the Base model run. In essence, we are looking for a relative measure of the differences, where sometimes the Base model's results will be closest to that observed and at other times another model run's results will be closest to that observed.

For this comparison, we inspect the distribution of 500 Diff-2 values for each of the other model runs, where Diff-2 = abs(observed-Base)-abs(observed-model). If the value of zero is encompassed between the 5th and the 95th percentile of the distribution of 500 Diff-2 differences for a model run, than we conclude that for this model run differences with the Base modeling results for this percentile are of no consequence.

2.5. Summarizing results from multiple monitors

The three comparisons outlined above are conducted at each site for each percentile in the cumulative frequency distribution of the maximum daily 8-h ozone values, namely: (1) which model run

is selected as Base; (2) whether the Base model's results compare favorably with that observed; and (3) whether the other modeling results are comparable or differ with the Base model run. We suggest as a summary statistic the number of times a model run is selected as the Base or is deemed to be performing as well as the Base model run for each percentile. For ease in later discussions, this summary statistic is referred to as the "Overall Skill score".

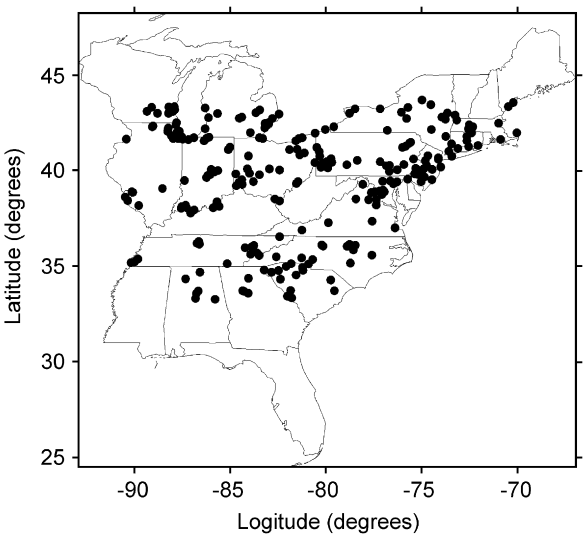

Fig. 1. The 242 monitoring locations that have sufficient observed hourly ozone values for analysis and also have modeling results from all four model simulations.

3. Modeling runs

Available for illustration of the type of results that can be obtained are ozone modeling results for the summer of 2002 from four regional-scale air quality simulations. These simulations were conducted using Version 4.5 of the Community Multi-Scale Air Quality Model (CMAQ) (Byun and Schere, 2006) by different modeling groups for various purposes. The modeling domains were somewhat different but there is a region in the central Eastern United States where we have ozone estimates for 242 monitoring sites from all models (see Fig. 1). There were differences in the setup of these model runs (see Table 1). For instance, the output grid size was 12-km for the first three model runs and 36-km for the fourth model run listed. Two of the simulations were conducted using identical processing of the meteorology, consistent processing of the emissions, and the same version of CMAQ. The only difference was the use of the CB-IV chemical mechanism (Gery et al., 1989) in one model run and the use of the SAPRC-99 chemical mechanism (Carter, 2000a, b) in the other model run. To develop a clearer assessment of relative modeling skill, we might desire comparison runs conducted with more similar treatments of meteorology and emissions (where possible) and with modeling results over a large extent of the United States. However, our focus here is to describe and illustrate a procedure for model inter-comparisons.

Table 1
Summary of setup configurations for model simulations

‘	Run 1	Run 2	Run 3	Run 4
Meteorology	MM5/FDDA	MM5/FDDA	MM5/FDDA	MM5/FDDA
Grid size (km)	12	36	12	12
Emissions	Ozone Transport Commission 2002 BaseB1	National Emission Inventory 2001 (no major point source CEM[a] emissions)	National Emission Inventory 2001 Grown (SMOKE[b] for vehicle emissions and CEM for major point sources)	National Emission Inventory 2001 Grown (SMOKE for vehicle emissions and CEM for major point sources)
Boundary conditions	GEOS-CHEM[c] CMAQ 36 km	Default	CMAQ 36 km	CMAQ 36 km
CMAQ version	4.5	4.5	4.5	4.5
Chemical mechanism	CB-IV	SAPRC-99	CB-IV	SAPRC-99
Reference	Ozone Transport Commission (2007)	Nolte et al. (2008)	Gilliland et al. (2008), Godowitch et al. (2008)	Gilliland et al. (2008), Godowitch et al. (2008)

[a]CEM: continuous emission monitoring (http://www.epa.gov/ttnnaaqs/ozone/areas/etscem.htm).
[b]SMOKE: Sparse Matrix Operator Kernel Emissions (SMOKE) Modeling System (http://www.smoke-model.org/index.cfm).
[c]GEOS-CHEM: global model of atmospheric composition driven by assimilated meteorological observations from the Goddard Earth Observing System (GEOS) (http://www-as.harvard.edu/chemistry/trop/geos/index.html).

4. Results and discussion

4.1. Example of results from a single site

Fig. 2 depicts the variation in the daily maximum 8-h ozone value over the 2002 summer season as observed at a monitoring site near Marion, KY (Site ID 050350005, latitude 35.197°N and longitude −90.191°E). The variations seen at this site are typical of that seen at other sites, as (1) there is some autocorrelation in the values from 1 day to the next (at this site the correlation coefficient, r, for a lag of 1 day is 0.68), and (2) there are large swings in the values resulting from synoptic-scale forcing, reflecting changes in air mass.

Fig. 3 depicts the results obtained from constructing 500 bootstrap summer seasons for the monitoring site near Marion, KY. For this analysis, we sampled 15 pairs of values from each month to create 90-day replicate seasons for analysis. From the 500 bootstrap summer seasons, we can derive 500 modeled and observed values for each percentile of interest. The bars in Fig. 3 illustrate the variability in the observed percentile values as estimated by the bootstrap sampling.

For the purposes of illustrating how the evaluation procedure works, we will show how results would be generated for comparisons of modeled and observed average 90th percentile values. Table 2 lists the averages, medians and standard deviations derived for the observed and modeled 90th percentile values. Also listed in Table 2 are the average and median of the absolute value of the difference between the modeled and observed percentile values, abs(Diff-1). From the results presented in Table 2 for this site near Marion, KY, the best performing model (Base model) is determined to be Run 4 for characterizing the 90th percentile of the cumulative frequency of maximum daily 8-h ozone values (smallest median value for abs(Diff-1)).

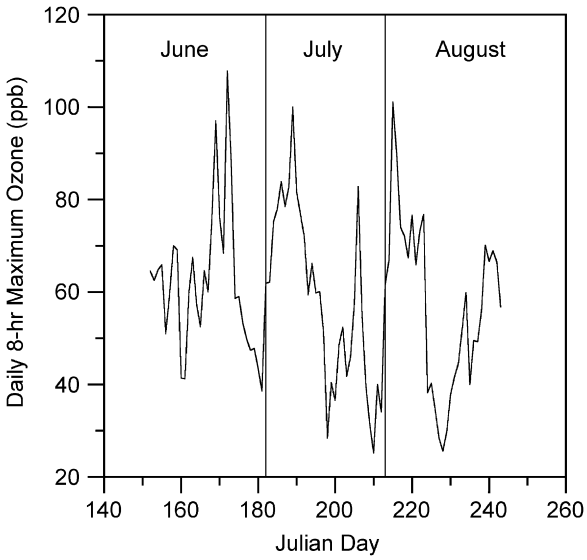

Fig. 2. Daily maximum 8-h ozone observed at a site near Marion, KY for the summer of 2002.

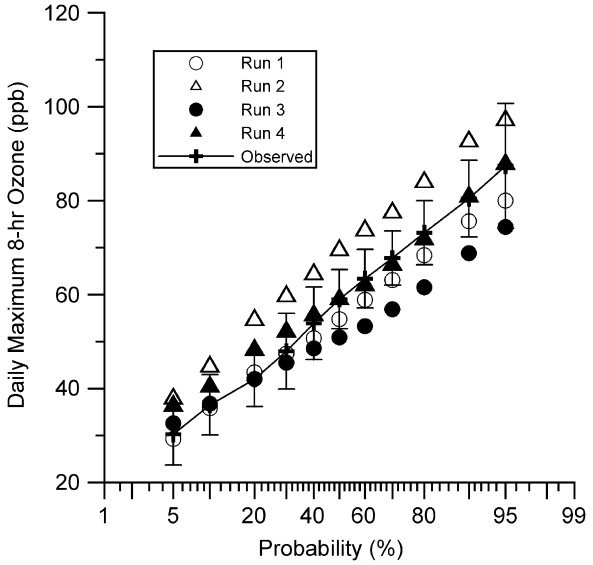

Fig. 3. Cumulative frequency plot of average daily maximum 8-h ozone observed and predicted derived from 500 bootstrap summer seasons at a site near Marion, KY for the summer of 2002. The error bars depict plus and minus two times the standard deviation about the computed observed average percentile values.

Fig. 4(A) illustrates the comparison for observed and modeled 90th percentile values of the distribution of the Base model's results with the observations (Diff-1 values). The value of zero is encompassed between the 5th and 95th percentiles of the distribution of Diff-1 values. It is concluded that Base model run's estimates of the average value for the 90th percentile value are similar to that observed.

Fig. 4(B) illustrates the comparison for observed and modeled 90th percentile values of the distribution of Diff-2 values, where we are testing to see if the results of the other model runs differ with the Base model's results. The value of zero is encompassed between the 5th and 95th percentiles of the distribution of Diff-2 values for Run 1 results, but such does not occur for the Run 2 or Run 3 results. It is concluded that Run 1 is performing as well as

Table 2
Summary of results for the 90th percentile in the cumulative frequency distribution for the site near Marion, KY

	Observed	Run 1	Run 2	Run 3	Run 4
Percentile from data	81.6	76.8	94.5	69.6	82.1
Average of boot replications	80.67	75.69	93.21	69.14	81.29
Standard deviation of boot replications	3.90	3.15	3.51	4.34	5.25
Median of boot replications	78.40	76.50	93.85	67.10	80.40
Average of abs(Diff-1)		5.05	12.55	11.53	2.69
Median of abs(Diff-1)		5.00	12.85	12.40	2.30

Listed are the 90th percentile values computed from the data and statistics of the 90th percentile derived from 500 bootstrap summers. All values are ozone concentration values (ppb).

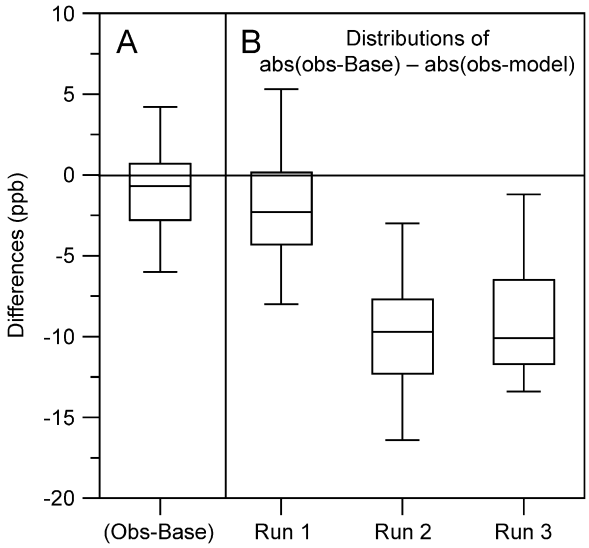

Fig. 4. Comparison of modeling results with observations for the 90th percentile value in the cumulative frequency of daily maximum 8-h ozone for the site near Marion, KY derived from 500 boot sample summer seasons. The line in each box is the median; the box depicts the inter-quartile range, and the extended bars depict the range of values from the 5th to the 95th percentile of the distribution of differences: (A) distribution of (Obs-Base) values; (B) distribution of Diff-2 = abs(observed-Base)−abs(observed-model) values.

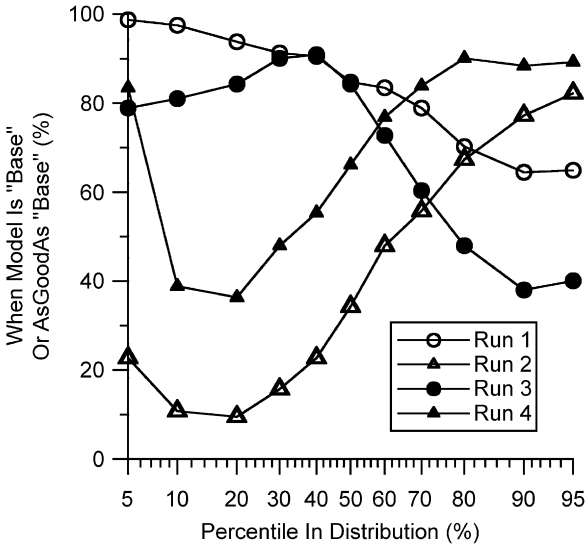

Fig. 5. Summary over 242 monitoring sites of which model is selected as Base or is deemed to be performing as well as the Base model run.

the Base model run, and Run 2 and Run 3 are not performing as well as the Base model run in predicting the average 90th percentile value.

4.2. Example of summarizing results across all monitors

Fig. 5 provides a quantitative summary of the "Overall Skill score" results obtained over all 242 monitoring locations. These results suggest Run 1 is performing as well or better than any other model run from the 5th through the 70th percentile values. These results are intriguing given that Run 1 and Run 3 are employing CB-IV; Run 2 and Run 4 are employing SAPRC-99, and the only difference between Run 3 and Run 4 is the choice of chemical mechanism. One obvious conclusion is that SAPRC-99 is producing higher ozone values than CB-IV when all other factors are equal, which confirms findings by Gilliland et al. (2008) and Sarwar et al. (2008). Given that different modeling groups constructed the emission inventories for Run 1 and Run 3, a diagnostic assessment of emissions might reveal answers for the differences seen, and whether certain emission characterization practices used by one group might be recommended more generally. However, it is beyond the scope of our discussion to explore these questions, as they represent major diagnostic research endeavors in themselves.

There are two questions yet to be resolved in defining the bootstrap sampling procedure: (1) Are the comparison results sensitive to autocorrelation in the ozone time series, and if so, how long of a block length is needed?; (2) How many bootstrap replications should be used in comparison analyses?

4.3. Sensitivity of results to block length

To investigate day-to-day (serial) correlation effects, an analysis of summer season 2002 ozone values for the 242 monitors with coverage over the Eastern United States was conducted. It was determined that the autocorrelation coefficient (r) of lag 1 (correlation between current day and previous day values) of daily maximum 8-h average daytime ozone values is on average over all sites approximately 0.52 (standard deviation of 0.07) from 1 day to the next and rapidly decreases as the lag separation in days increases (Fig. 6). The values of r ranged from a minimum of 0.23 to a maximum of 0.73, and we detected a slight correlation with latitude (correlation coefficient, r^2, of 0.2), being largest for the southern states and decreasing as latitude increases.

To maintain the approximate effects of day-to-day correlation structure in the observations when bootstrap sampling a month, we can select short blocks (e.g., 2 consecutive days-pairs, 3 consecutive days-triplets) of values rather than selecting independent samples with replacement. A formal analysis of the lagged correlations could be used to select the length of the block (Wilks, 1997). We decided to employ a more *ad hoc* approach, in which we investigate the sensitivity of our final results to various choices of block length. To investigate this, three sensitivity runs were made, in which the sample block length increased successively from sampling independent days (block length of one), to pairs of days, to triplets of days (block length of three) from each month. These analyses were conducted using 500 bootstrap replicates. For each percentile from the 5th through to the 95th, we tracked the change in the "Overall Skill score", namely the number of times a model run is selected as the Base or is deemed to be performing as well as the Base model run, as the sample block length increased. If the comparison results are sensitive to autocorrelation effects, the Overall Skill score for each model should increase as sample block length increases. This should occur because if the autocorrelation in the time series is sufficient, the variance of sample estimates of the percentile values would increase as the sample block length increases, and larger variances of the sample estimates of the percentile values would make it more likely for a model run to be deemed to be performing as well as the Base model run (all other factors being equal).

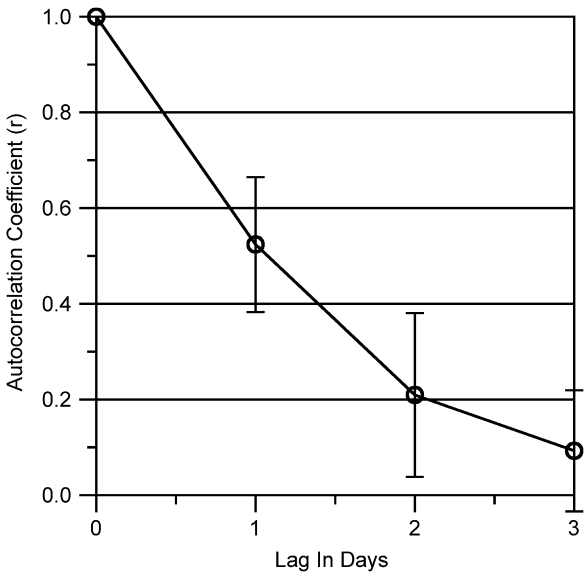

Fig. 6. Correlation in daily maximum 8-h ozone values from 1 day to the next (lag of 1 day), and out to a lag of 3 days. Averages are computed over 242 sites and error bars depict plus and minus two times the standard deviation about the computed averages.

Inspection of the results revealed that as the block length increases there was a small incremental increase in the Overall Skill score for all of the models as the block length increased. The increase in the Overall Skill score (averaged over all models and percentiles) as the block length increased from a sample block length of one to a sample block length of two was 0.86%. The increase in the Overall Skill score (averaged over all models and percentiles) was 1.10% in going from a sample block length of one to a sample block length of three, which shows that most of the autocorrelation effects are addressed with a sample block length of two.

4.4. Sensitivity of results to number of bootstrap replications

Fig. 7 summarizes a sensitivity test to assess the variability of the summary results to the number of bootstrap replications (summer seasons) generated for analysis. For this analysis the evaluation procedure was run 25 times (with different random

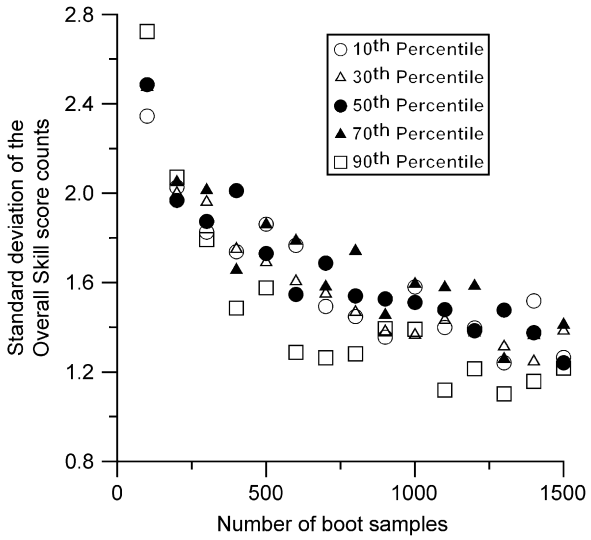

Fig. 7. Standard deviation of the "Overall Skill score" for the four models runs as a function of sample boot size of summer seasons generated for analysis. Results are shown for the 10th, 30th, 50th, 70th and 90th percentiles in the cumulative frequency of daily maximum 8-h ozone values. The "Overall Skill score" is the number of times a model run is selected as the Base or is deemed to be performing as well as the Base model run over the 242 monitoring sites.

seed start values) for each percentile with a predefined boot replication size. The standard deviation in the Overall Skill score for each model was computed from the 25 runs for each percentile. These 25 runs and subsequent computations were then conducted for increasingly larger boot replication sizes. As seen in Fig. 7, the variability in the Overall Skill score counts for each percentile decreases as the number of boot replications increases. If would appear that having bootstrap replications on the order of 500–750 reduces the variability in the summary counts to on the order ± 2 or less. Larger sample sizes can reduce the variability somewhat, but such reductions are likely of no practical consequence.

5. Summary

An operational model inter-comparison procedure is described that quantitatively assesses the relative skill among several regional-scale air quality model simulations to predict various percentiles in the cumulative frequency distribution of observed daily maximum 8-h ozone concentration values. The procedure determines at each monitoring site for each percentile in the cumulative frequency distribution which one of several model runs is performing best, called the Base model run, and then determines whether the Base model run's prediction differ with the observed value, and whether the results for the other model runs differ with the Base model run. In our discussion, we illustrated the inter-comparison procedure using daily maximum 8-h ozone values. A pragmatic feature of the inter-comparison procedure is that it requires a minimum of data to be shared between modeling groups, namely a listing of each model run's predictions of daily maximum 8-h average ozone values for a summer season. The procedure would work as well for daily maximum 1-h ozone values. A future extension of this procedure might include a definition of "best performing model" as that model run that routinely is closest to several chemical species deemed of importance simultaneously in comparison to other model runs. Another future extension of this procedure might investigate methods for assessing performance for the more extreme percentile values (e.g., Rao et al., 1985).

The variability in the observed cumulative frequency distribution of daily maximum 8-h ozone values is estimated using bootstrap sampling. The sampling is conducted in a manner to capture autocorrelation inherent in the observations, and effects of model-to-observations correlations and model-to-model correlations. It was determined that suitable comparison results can be obtained through the use of bootstrap sampling of pairs of days and with 500–750 bootstrap replications of pseudo summer seasons.

The daily maximum 8-h average ozone values are estimated by a complex modeling system that attempts to characterize emissions, meteorology, transport, chemical boundary conditions, and fate of the emissions. Operational inter-comparisons as described here cannot provide information regarding whether "better correspondence with observations" is due to better science or due to offsetting biases. Operational model evaluations are not a substitute for diagnostic model evaluations, but operational inter-comparison of several models can provide information on whether model results are comparable in an overall sense, and perhaps may provide ideas where diagnostic evaluations might prove insightful. It is our hope that this discussion will encourage the development of standardized benchmark calculations and promote systematic records of progress of air quality modeling.

Disclaimer: The research presented here was performed under the Memorandum of Understanding

between the US Environmental Protection Agency (EPA) and the US Department of Commerce's National Oceanic and Atmospheric Administration (NOAA) and under agreement number DW 13921548. John S. Irwin collaborated in the research described here under contract EPA Order No. EP06D000631. This work constitutes a contribution to the NOAA Air Quality Program. Although it has been reviewed by EPA and NOAA and approved for publication, it does not necessarily reflect their policies or views.

References

ASTM International, 2000. Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance (D 6589), (available at 〈http://www.astm.org〉), 100 Barr Harbor Drive, P.O. Box C700, West Conshohocken, PA 19428, USA, 2000 (updated 2005), p. 17.

Biswas, J., Hogrefe, C., Rao, S.T., Hao, W., Sistla, G., 2001. Evaluating the performance of regional-scale photochemical modeling systems. Part III—Precursor predictions. Atmospheric Environment 35, 6129–6149.

Byun, D.W., Schere, K.L., 2006. Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. Applied Mechanical Reviews 55, 51–77.

Carter, W.P.L., 2000a. Documentation of the SAPRC-99 chemical mechanism for VOC reactivity assessment. Draft Report to the California Air Resources Board, Contracts 92-329 and 95-308. Available at 〈http://www.cert.ucr.edu/~carter/absts.htm#saprc99〉 (accessed 8 May).

Carter, W.P.L., 2000b. Implementation of the SAPRC-99 chemical mechanism into the Models-3 framework. Report to the United States Environmental Protection Agency. Available at 〈http://www.cert.ucr.edu/~carter/absts.htm#s99mod3〉 (accessed 29 January).

Dabberdt, W., Carroll, M., Baumgardner, D., Carmichael, G., Cohen, R., Dye, T., Ellis, J., Grell, G., Grimmond, S., Hanna, S., Irwin, J., Lamb, B., Madronich, S., McQueen, J., Meagher, J., Odman, T., Pleim, J., Schmid, H.P., Westphal, D., 2004. Meteorological research needs for improved air quality forecasting. Bulletin of the American Meteorological Society 85 (4), 563–586.

Demerjian, K.L., 1985. Quantifying uncertainty in long-range-transport models: a summary of the AMS workshop on sources and evaluation of uncertainty in long-range-transport models. Bulletin of the American Meteorological Society 66, 1533–1540.

Fox, D.G., 1981. Judging air quality model performance: a summary of the AMS workshop on dispersion model performance. Bulletin of the American Meteorological Society 62, 599–609.

Fox, D.G., 1984. Uncertainty in air quality modeling. Bulletin of the American Meteorological Society 65 (1), 27–36.

Gery, M., Whitten, G., Killus, J., Dodge, M., 1989. A photochemical kinetics mechanism for urban and regional scale computer modeling. Journal of Geophysical Research 94, 925–956.

Gilliland, A.B., Hogrefe, C., Pinder, R.W., Godowitch, J.L., Foley, K.M., Rao, S.T., 2008. Dynamic evaluation of regional air quality models: assessing changes in ozone stemming from changes in emissions and meteorology. Atmospheric Environment, in press.

Godowitch, J.L., Hogrefe, C., Rao, S.T., 2008. Influence of point source NO_x emission reductions on modeled processes governing ozone concentrations and chemical/transport indicators. Journal of Geophysical Research, in review.

Hanna, S.R., Gifford, F.A., 1971. Summary of meeting on mesoscale atmospheric modeling. Bulletin of the American Meteorological Society 52, 993.

Hogrefe, C., Rao, S.T., Kasibhatla, P., Kallos, G., Tremback, C., Hao, W., Olerud, D., Xiu, A., McHenry, J., Alapaty, K., 2001a. Evaluating the performance or regional-scale photochemical modeling systems: Part I—Meteorological predictions. Atmospheric Environment 35, 4159–4174.

Hogrefe, C., Rao, S.T., Kasibhatla, P., Kallos, G., Hao, W., Sistla, G., Mathur, R., McHenry, J., 2001b. Evaluating the performance or regional-scale photochemical modeling systems: Part II—Ozone predictions. Atmospheric Environment 35, 4175–4188.

Hogrefe, C., Porter, P.S., Gego, E., Gilliland, A., Gilliam, R., Saqll, J., Irwin, J., Rao, S.T., 2006. Temporal features in observed and simulated meteorology and air quality over the Eastern United States. Atmospheric Environment 40, 5041–5055.

Julian, P.R., Murphy, A.H., 1972. Probability and statistics in meteorology: a review of some recent developments. Bulletin of the American Meteorological Society 53, 957–965.

Lorenz, E., 1963. Deterministic non-periodic flow. Journal of Atmospheric Science 20, 130–141.

Lorenz, E., 1983. A history of prevailing ideas about the general circulation of the atmosphere. Bulletin of the American Meteorological Society 64, 730–769.

Nolte, C.G., Gilliland, A.B., Hogrefe, C., Mickley, L.J., 2008. Linking global to regional models to assess future climate impacts on surface ozone levels in the United States. Journal of Geophysical Research, in press.

Ozone Transport Commission, 2007. Modeling: technical support documentation for use in state implementation plan development. 〈http://www.otcair.org/whats_new_details.asp?PID=61〉.

Rao, S.T., Sistla, G., Pagnotti, V., Petersen, W.B., Irwin, J.S., Turner, D.B., 1985. Resampling and extreme value statistics in air quality model performance evaluation. Atmospheric Environment 19, 1503–1518.

Sarwar, G., Luecken, D., Yarwood, G., Whitten, G., Carter, B., 2008. Impact of an updated carbon bond mechanism on air quality using the community multiscale air quality modeling system: preliminary assessment. Journals of Applied Meterorology and Climatology 47, 3–14.

Venkatram, A., 1979. The expected deviation of observed concentrations from predicted ensemble means. Atmospheric Environment 13, 1547–1549.

Venkatram, A., 1983. Uncertainty in predictions from air quality models. Boundary Layer Meteorology 27, 186–196.

Weil, J.C., Sykes, R.I., Venkatram, A., 1992. Evaluating air quality models: review and outlook. Journal of Applied Meteorology 31 (10), 1121–1145.

Wilks, D., 1997. Sampling hypothesis tests for autocorrelated fields. Journal of Climate 10, 65–82.

Yoden, S., 2007. Atmospheric predictability. Journal of Meteorological Society of Japan, Series II 85B, 77–102.