AN OPERATIONAL MODEL EVALUATION PROCEDURE FOR ASSESSING THE RELATIVE SKILL BETWEEN COMPETING AIR QUALITY MODELS IN ESTIMATING 8-HOUR MAXIMUM OZONE VALUES

John S. Irwin
John S. Irwin and Associates, Raleigh, NC

INTRODUCTION

The ASTM Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance (D 6589, available at: http://www.astm.org/) recommends three principles to be followed in designing operational model evaluation methods for air quality models:

1) Model performance (or "skill") has meaning only through comparison with existing competition; thus to determine model performance requires direct comparisons with competing models.

2) Air quality models predict what is to be seen on average and are not capable of replicating short-term or small-scale variations in the observations, thus comparisons of modelling results and observations should be conducted using a well-defined spatial or temporal average of some feature in the observed concentrations.

3) The design of the model evaluation method should provide a quantitative test of whether differences seen between the best performing model and its competition are statistically significant.

4) A fourth principle (not stated in ASTM D 6589) is that when differences are deemed statistically significant, information should be provided that allows a qualitative assessment of whether these differences are of practical concern.

This presentation outlines the steps taken to develop an operational model evaluation method that meets the four objectives listed above, and that assesses in a quantitative manner the relative skill among several competing air quality models to replicate the observed average 8-hour maximum ozone.

DISCUSSION

Modelling results

Ozone modelling results are available for the summer of 2002 from four (4) air quality models: NOXSIP-CB4, NOXSIP-SAPRC, OTCBaseB1, Chesapeake. These runs were conducted at different times for different uses, using the Community Multi-scale Air Quality Model (CMAQ). There are differences in the setup and conduct of these model runs. For instance, the output grid size was 12-km for the first three models and 36-km for the fourth model listed. The modelling domains were somewhat different but there is a region in the central Eastern United States where we have ozone estimates for 248 monitoring sites from all models. To develop a clearer assessment of relative modelling skill, we might desire runs conducted with more similar treatments of meteorology and emissions (where possible) with modelling results over a larger extent of the United States for comparison. However, our purpose in this presentation is to describe the comparison procedure and to stimulate a discussion on how the comparison procedure might be improved.

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 CB4 chemical mechanism (NOXSIP-CB4) in one model run and the use of the SAPRC chemical mechanism (NOXSIP-SAPRC) in the other model run. It was anticipated that detecting differences in skill between these two model runs might prove difficult. Hence, we used these two model runs in our preliminary test runs to assess how best to design the model evaluation procedure.

Defining the observed feature (8-hour maximum ozone)

The observations contain variations that are impossible to replicate by air quality models, in part because of lack of understanding of all the physical processes; in part because of uncertainties in the model inputs, and in part because a portion of the observed variations is stochastic and thus cannot be deterministically simulated.

Various investigations have shown that current regional-scale air quality models have skill in replicating the synoptic time-scale variations in the observations, but lack skill in replicating short-term (hourly) variations in the observations. Therefore to place the comparisons within the skill range of current models, the feature we will challenge the models to replicate is the average 8-hour maximum ozone. This can be defined two ways: 1) by computing hourly ozone values averaged over all days in the month and then determining the 8-hour maximum from these hourly ozone concentration averages, and 2) by computing the 8-hour maximum ozone for each day of the month and then averaging all these values for the month. These likely provide a similar test for model performance, so we will try each one, anticipating that the second definition may ultimately prove to be faster in computation time.

Defining the bootstrap resampling procedure

Since the observations represent a sample from a population of possible outcomes, features derived from the available set of observations (e.g., 8-hour maximum) are estimates that have uncertainty since we do not have access to the entire set of possible outcomes. One means to estimate the uncertainty is to employ bootstrap resampling. Bootstrap resampling involves building pseudo-sets of observations by sampling the available set of observations with replacement. A goal in the design of the resampling procedure is to maintain correlations that might exist between the observations in each of the pseudo-sets of observations. Correlations in a time series of values, affects the variance of the time series, making it larger, all other factors being equal. A larger variance in the observations would increase the uncertainty (variance) in features derived from the pseudo-sets of observations. A larger variance in derived features means that differences between what is modelled and that observed would need to be larger before such differences would be deemed statistically significant.

An analysis of July 2002 ozone values for 760 monitors with coverage over the Eastern half of the United States was conducted. It was determined that the correlation coefficient (r^2) between daytime ozone values was approximately 0.85. It was also determined that the correlation coefficient from one day to the next varied by the hour of the day, as depicted in Figure 1, being strongest during mid-day with a correlation coefficient of approximately 0.15. To maintain the hour-to-hour correlations, we sample whole days in developing our pseudo-months of observations for analysis. To maintain the day-to-day correlations, we can sample pairs of days. We will try each way to see if the results are affected.

A goal in the resampling procedure is to maintain correlations that might exist model-to-model and model-to-observations. This can be attained by concurrently sampling the modelling results associated with the respective days of observations selected in constructing pseudo-sets of observations for analysis.

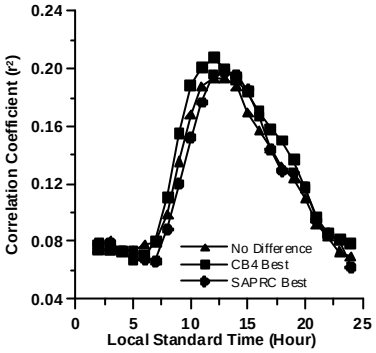

Fig. 1. Diurnal variation in correlation coefficient (r^2) from one day to the next.

Statistical test to detect differences in skill

We conduct our assessment of skill independently at each monitor. The metric to assess skill will be how close each model's prediction of the month-average 8-hour maximum ozone value is to that observed. The actual metric is the absolute value of the difference of the observed value minus the modelled value, Diff = Abs(Obs-Est). In the preliminary runs used to complete design considerations, we attempted to detect differences in skill between results by NOXSIP-CB4 and NOXSIP-SAPRC. In this simple two-way assessment, for each boot sample (pseudo-set) we compute Diff#1 = Abs(obs-CB4) and Diff#2 = Abs(obs-SAPRC), where obs, CB4 and SAPRC represent the average 8-hour maximum ozone values (observed and modelled) derived from the boot sample.

For each boot sample, we compute the difference of the differences, DIFF = Diff#1-Diff#2. At the end of the boot sampling, we compute the average and standard deviation of DIFF, and compute a t-Test = Std(DIFF)/Avg(DIFF). If the t-Test is less than -1.96, then we conclude the CB4 model estimates of the average 8-hour ozone maximum are closer to that observed. If the t-Test is greater than +1.96, then we conclude the SAPRC model estimates of the average 8-hour ozone maximum are closer to that observed. If the t-Test is between -1.96 and +1.96, we conclude we can not detect any difference in the modelling results at this monitor.

Table 1. July 2002 comparison results using month-averaged hourly ozone values. A comparison is provided of the differences to be seen in results attained when pseudo-sets of data are constructed using pairs of days versus individual days. Listed in the table is the number of monitors at which skill differences were or were not detected.

	Individual Days	Pairs of Days
NOXSIP-CB4	270	256
No differences in skill	352	377
NOXSIP-SAPRC	138	128

Results of sensitivity tests

In our first implementation of the comparison procedure, we derived the average 8-hour maximum ozone from month-averaged hourly values of ozone. Once this was working, we

tested to see if the conclusions reached changed if we constructed our boot samples using pairs of days rather than individual days. There were subtle but consistent differences in the results (see Table 1) suggesting that the day-to-day correlations picked up by sampling pairs of days did alter the results in the expected manner. That is, when pairs of days were selected in the boot sampling, we detected differences in performance at fewer monitors than when individual days were selected.

In our second implementation of the comparison procedure, we computed the 8-hour maximum ozone for each day of July 2002 and then computed the average 8-hour maximum ozone by averaging the daily values. Once this was working, we tested to see if the conclusions reached were similar to those listed in Table 1 which was attained using month-averaged hourly ozone values. The results of these tests are listed in Table 2, and through comparison can be seen to be similar.

Table 2. July 2002 comparison results using daily 8-hour maximum ozone values. A comparison is provided of the differences to be seen in results attained when pseudo-sets of data are constructed using pairs of days versus individual days. Listed in the table is the number of monitors at which skill differences were or were not detected.

	Individual Days	Pairs of Days
NOXSIP-CB4	265	254
No differences in skill	355	373
NOXSIP-SAPRC	140	133

The results shown in Tables 1 and 2 were developed using 500 boot samples at each monitor. There was a substantial difference in the run times, where it took a bit over two hours to develop the results shown in Table 1 and it took less than 5 minutes to develop the results shown in Table 2.

From the sensitivity tests thus far summarized, we concluded that 1) suitable comparison results can be obtained using daily 8-hour maximum ozone values, and 2) sampling by pairs of days is warranted, as the day-to-day correlations are sufficient to affect results.

Our next sensitivity test was to see how sensitive the results were to the number of boot samples used in developing the results. To assess the variability of the results, twenty-five (25) runs were made at each monitor using different seed values for the numerical random sampling procedures. This provided us with an average and a standard deviation of the results listed in Table 2 using pairs of days sampling. This variability test was conducted using boot samples ranging from 125 to 4000 samples. The results of this test are listed in Table 3. It would appear that using boot samples of 500 to 1000 are warranted.

Practical significance of differences in skill detected

The results summarized have involved the comparison of two similar model runs of CMAQ in which one run employed the CB4 chemical mechanism and the other run employed the SAPRC chemical mechanism. From an analysis of the daily 8-hour maximum ozone values for July 2002, it was determined that the SAPRC maximum average 8-hour ozone values are 15% greater (on average) than that estimated using CB4.

At about half of the monitors, the estimates by both models were deemed to be similar in skill. CB4 was deemed to provide average maximum 8-hour ozone values closer to the observed values than SAPRC at 66% of the remaining monitors. Since there is a consistent

15% difference (on average) between the estimates by the two models, if the estimates by NOXSIP-CB4 are (on average) within 10% of that observed, then the NOXSIP-SAPRC results are deemed to be significantly different, and visa versa.

Table 3. Sensitivity of sampling results to number of boot samples used in the comparisons. Listed are the number of monitors at which skill differences were or were not detected. The averages and standard deviations (Stdev) were computed from 25 samples at each monitor holding the number of boot samples constant.

Number of Boot samples	CB4 Average	CB4 Stdev	No Difference Average	No Difference Stdev	SAPRC Average	SAPRC Stdev
125	254.80	1.62	371.84	2.80	133.36	2.08
250	253.68	1.44	372.96	2.13	133.36	1.74
500	254.00	1.10	372.68	1.72	133.32	1.49
1000	254.24	1.39	372.12	1.63	133.64	0.79
2000	254.00	1.20	372.56	1.39	133.44	0.70
4000	254.60	1.20	371.72	1.43	133.68	0.79

The comparison procedure appears to be capable of detecting consistent differences of at least 15%, which could be of practical concern. For instance, even though current regional-scale models are incapable of replicating localized extremes (maxima or minima), they are often used for this purpose. If one were to compare NOXSIP-CB4 and NOXSIP-SAPRC results to replicate the actual daily 8-hour maximum ozone values, "good comparisons" of extreme values by NOXSIP-SAPRC could fortuitously arise because the SAPRC chemical mechanism generates higher ozone values than CB4. But is this a result of better science or is this a result that SAPRC has a bias to estimate higher ozone values than CB4? If the favourable results using SAPRC are not the result of better science, will subsequent regulatory analyses of controls needing to be applied to emissions be valid or defensible?

SUMMARY

It has been shown that operational model evaluation procedures can detect differences in skill between models to replicate features in the observations, but they do not 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 they can provide ideas where diagnostic evaluations might prove insightful.

It was determined that suitable comparison results could be obtained through the use of daily 8-hour maximum ozone values; bootstrap resampling of pairs of days, and with bootstrap samples of 500 to 1000.

In this presentation, we followed the principles set forth in ASTM D 6589, which recommends testing air quality models in their ability to replicate average time or space variations, and to not test them on their ability to replicate small-scale or short-term extremes (for which none of the models have the requisite science). This is an opinion that may not be shared by a majority in the modelling community, but it is a topic that deserves debate and needs to be resolved as it has direct impact on the development of model evaluation methods.

More to come: Thus far a summary has been provided of the sensitivity test results developed to finalize details of an operational model evaluation procedure. Now that these design questions have been addressed, the next step is to demonstration the results obtained when we compare results obtained from four (4) models …