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

By

John S. Irwin
John S. Irwin and Associates, Raleigh, NC, 27615

11th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes
2-5 July 2007
Cambridge, UK

# ABSTRACT

This presentation outlines the status of an effort to develop an operational model evaluation method that assesses in a quantitative manner the relative skill among several competing air quality models to replicate the observed daily 8-hour maximum ozone concentration.

As a basic philosophy, we have attempted to follow the principles outlined in ASTM Standard Guide D6589, entitled, Statistical Evaluation of Atmospheric Dispersion Model Performance, (see next slide).

# Principles To Follow For Model Evaluation Methods

1) Defining a model's skill (i.e., How well is this model doing?) has meaning 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 modeling 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.

Ozone modeling results were available for the summer of 2002 from four (4) air quality model simulations: B1, Chesapk, CB4, and SAPRC

These runs were conducted using the Community Multi-scale Air Quality (CMAQ) Model. There are differences in the set-up and input data for these model runs. All simulations employed MM5/FDDA and CMAQ (version 4.5).

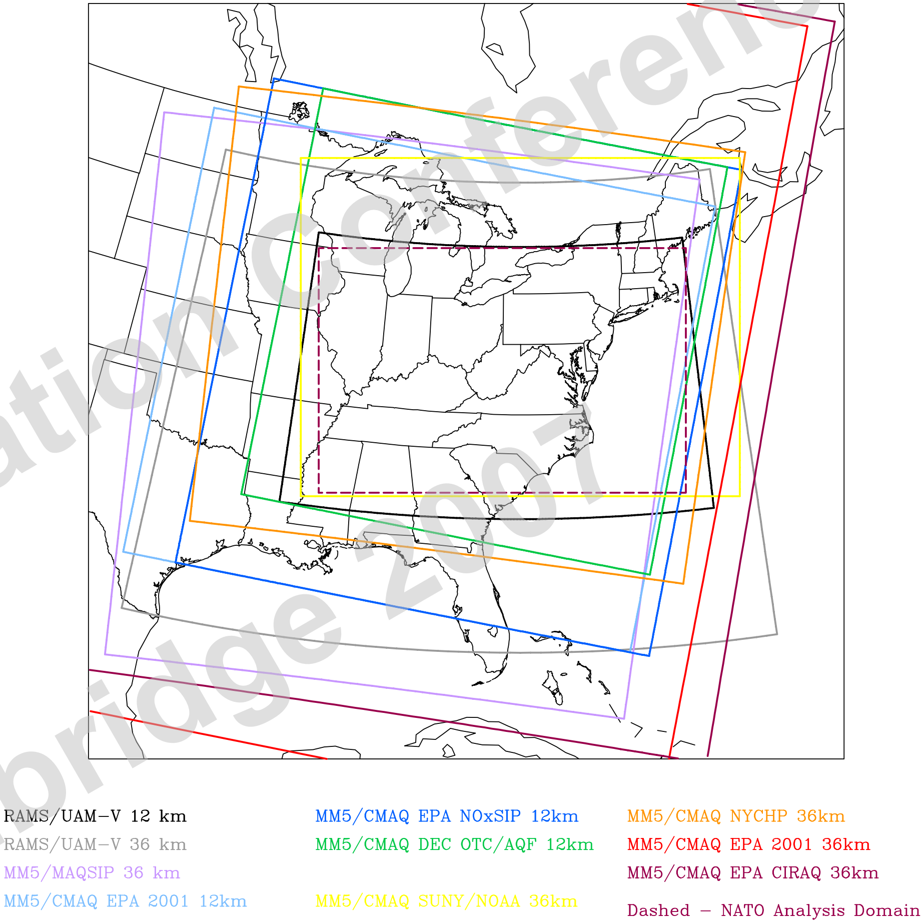
| **Abbrevia-tion** | **Grid Spacing (km)** | **Chem. Mecha-nism** | **Emissions** | **Reference** |
|---|---|---|---|---|
| **B1** | 12 | CB4 | OTC BaseB1 2002 | Ozone Transport Commission, 2007 |
| **Chesapk** | 36 | SAPRC | NEI2001 | Nolte et al., 2007 |
| **CB4** | 12 | CB4 | NEI 2002 | Gilliland et al. 2007<br>Godowitch et al. 2007 |
| **SAPRC** | 12 | SAPRC | NEI 2002 | Gilliland et al. 2007<br>Godowitch et al. 2007 |

# Analysis Domain

- Selected an analysis domain that was covered by all modeling grids (over land).

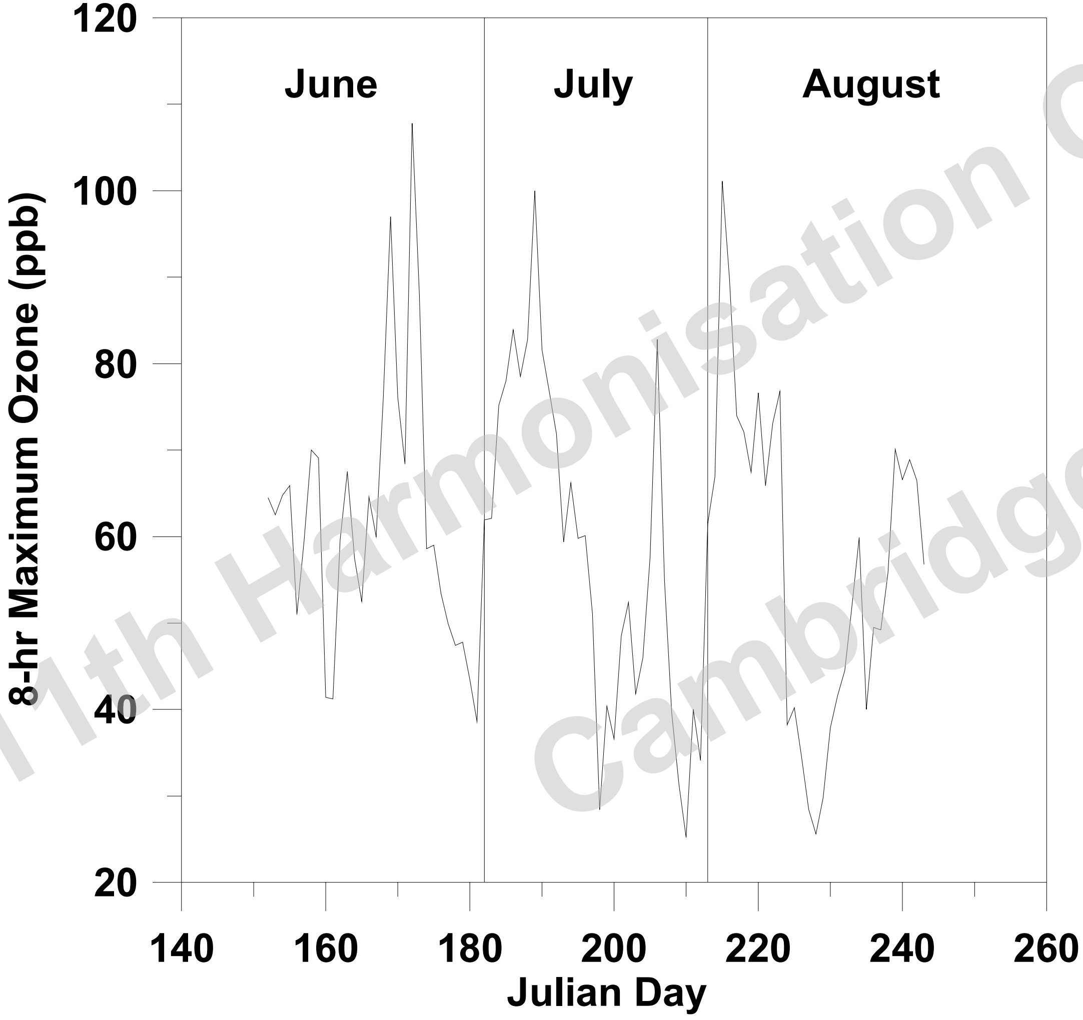
- Extracted hourly surface ozone observations from AQS at sites that had at least 50% non-missing data for summer season (June – August).

- This resulted in 248 sites in the analysis domain of which 242 had sufficient data for analysis for 2002.



- To match model values to observations, the model grid cell containing the monitoring location was used
- Data extraction and compilation was originally conducted for 11 model simulations listed, but there are only 4 simulations available for 2002.

Average daily maximum 8-hr ozone concentration observed at a site near Marion, Kentucky (Site 050350005) for the summer of 2002



We expect to see differences between that which is observed and that simulated, because the models only simulate a portion of the variations to be seen (Principle #2).

The observations represent what is seen at a particular point, whereas the models provide volume-averages.

We expect regional-scale models to properly characterize the seasonal- and synoptic-scale variations.

Fine-scale and short-lived variations cannot be properly characterized by the regional-scale-models.

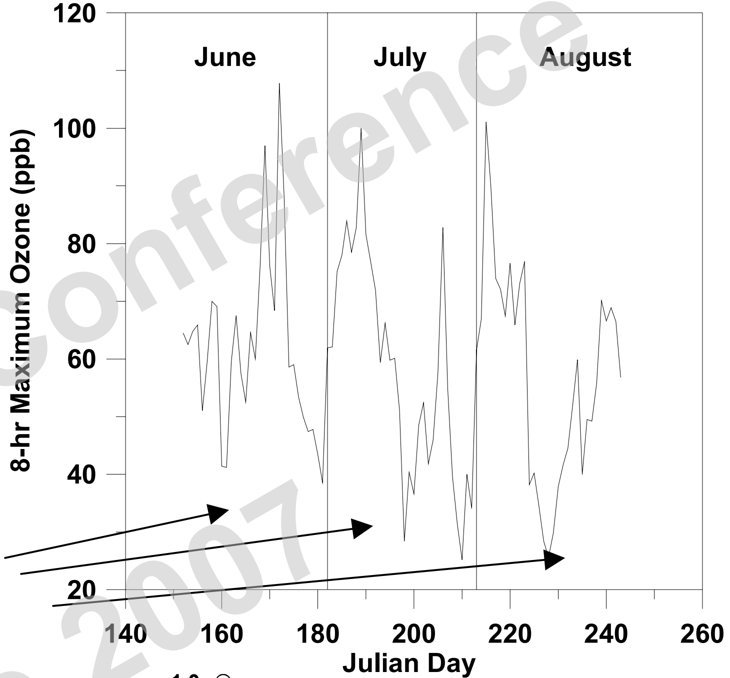
# Purpose For Bootstrap Resampling

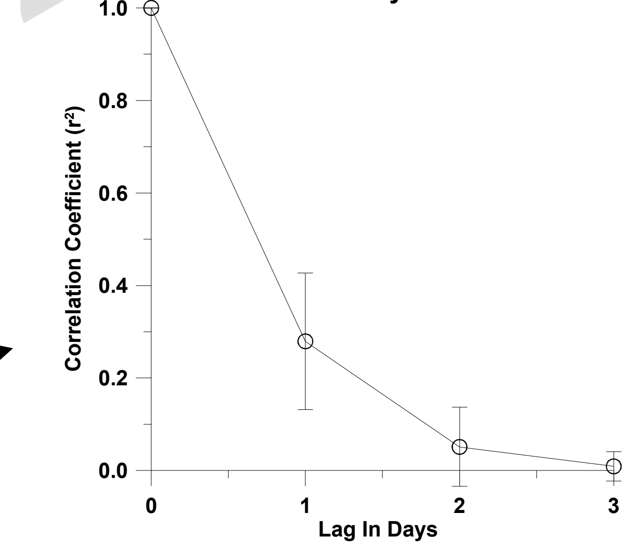
At best, we typically have a model simulation for a summer season, when what we need to place confidence bounds on model evaluation results (Principle #3) are multi-year model simulations.

We sample 30 values from each month to capture any month-to-month “seasonal trends.”

By resampling each month’s results with replacement, we are simulating the effects of experiencing a different collection of synoptic events.

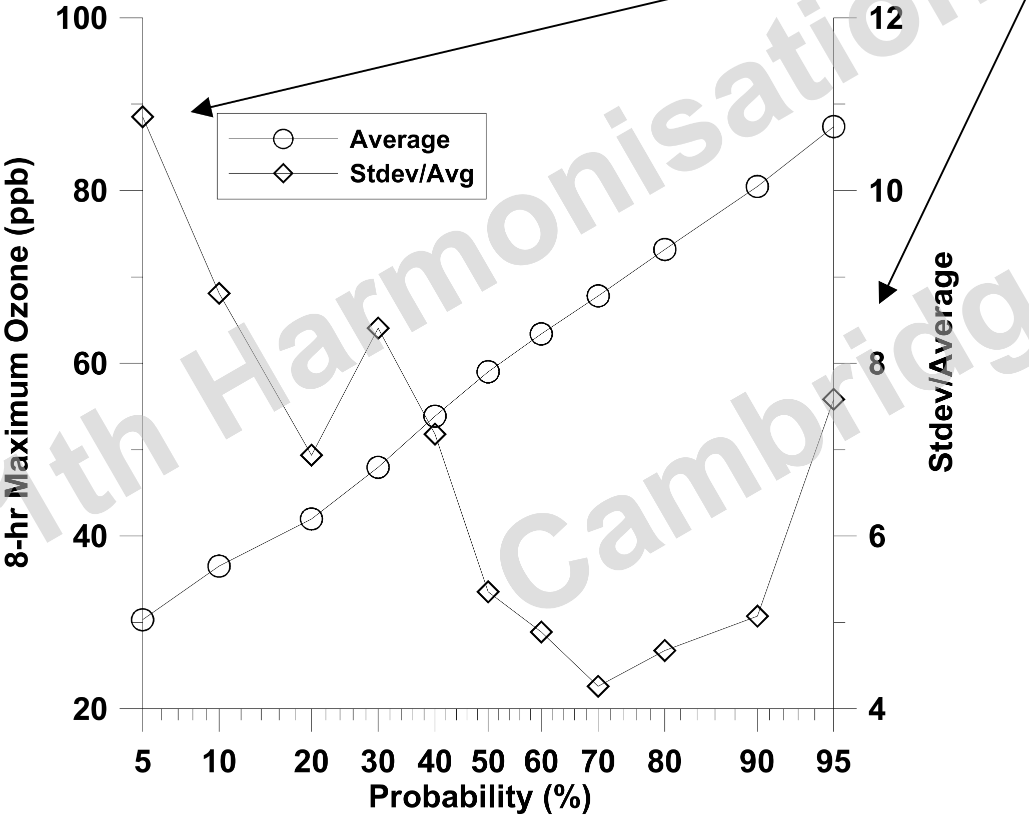
Draw pairs of values to capture the effects of “synoptic events” which induce strong correlations over 2-day periods.





# Results From Bootstrap Resampling The Observed Ozone Values

Analysis of 500 “seasons” of daily maximum 8-hr maximum ozone observed at a site near Marion, Kentucky (Site 050350005) for the Summer of 2002

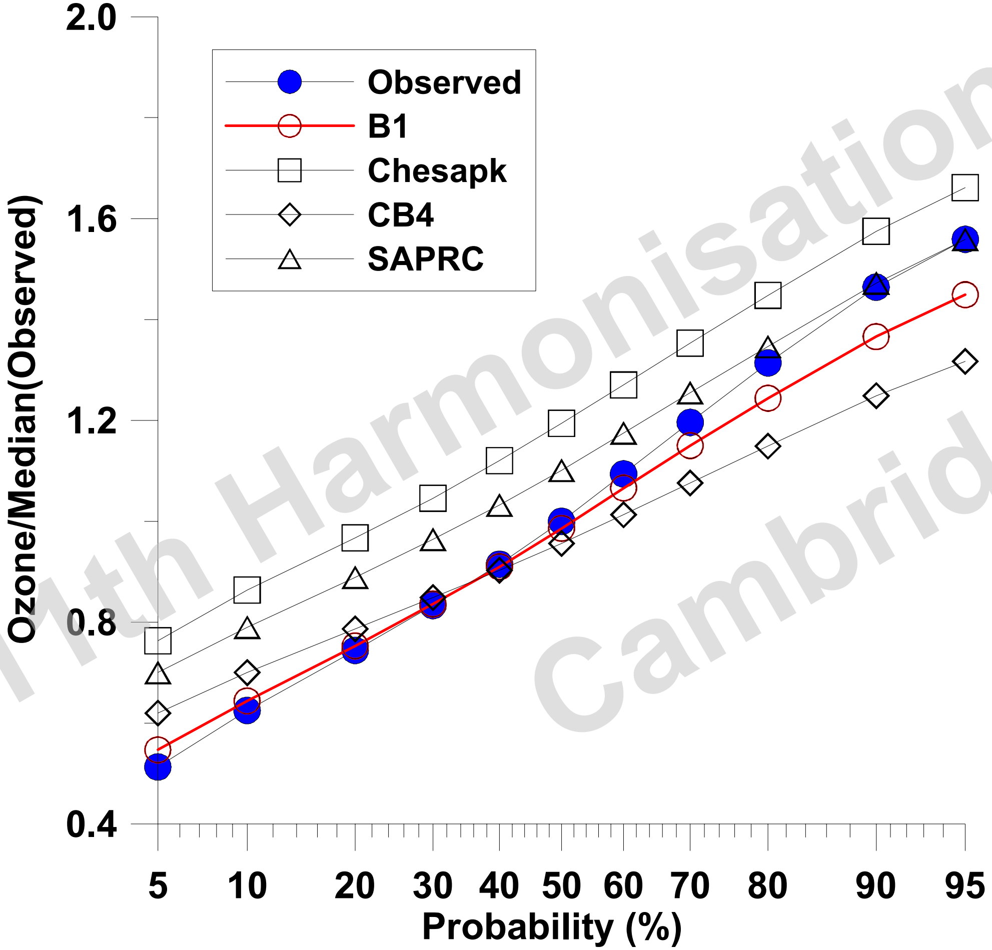


## The Cumulative Frequency Plot of Observed 8-hr Ozone Averages.

We expect that the relative variation for the more extreme percentile values will be larger than that seen for the middle percentile values.

This is expected because the middle percentile values are dominated by seasonal and synoptic variations, whereas the extreme percentile values reflect rarely occurring (and likely) random events (Principle #2).

**Cumulative Frequency Distribution of Normalized Observed and Simulated Maximum 8-hr Values Averaged Over All 242 Sites.** (At each site the average percentile values (determined by bootstrap resampling) were normalized by dividing by the median observed ozone value at each site)



Notice that the B1 results (red circles) appear to correspond best with the observations (blue circles), except for the highest percentile values.

For the upper percentile values, the SAPRC results (triangles) are in closest correspondence with the observations.

The “relative” skill varies as a function of percentile (Principle #1).

Can we characterize the relative skill in a quantitative manner, and discern whether differences are really significant (Principles #1 and #3)?

## Development Of Model Evaluation Comparison Statistics

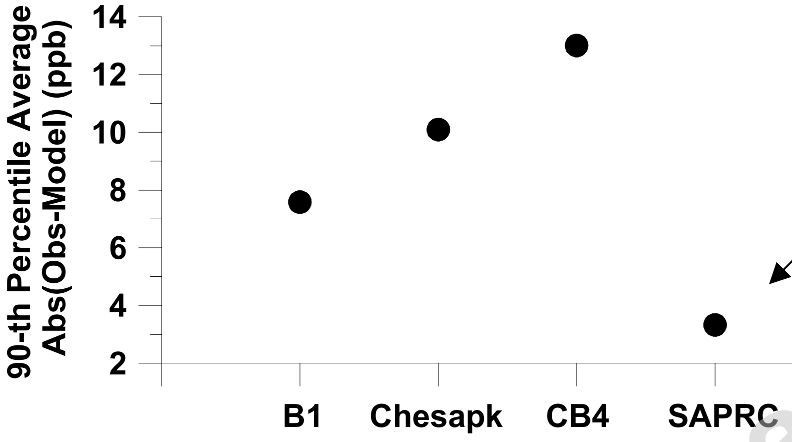
-When we select a pair of observed daily maximum 8-hr ozone values, we also pull the associated modeled values. This maintains obs-to-model and model-to-model biases.

-We sort each “Summer Season” of observed ozone values (smallest to highest) carrying along the associated modeled values.

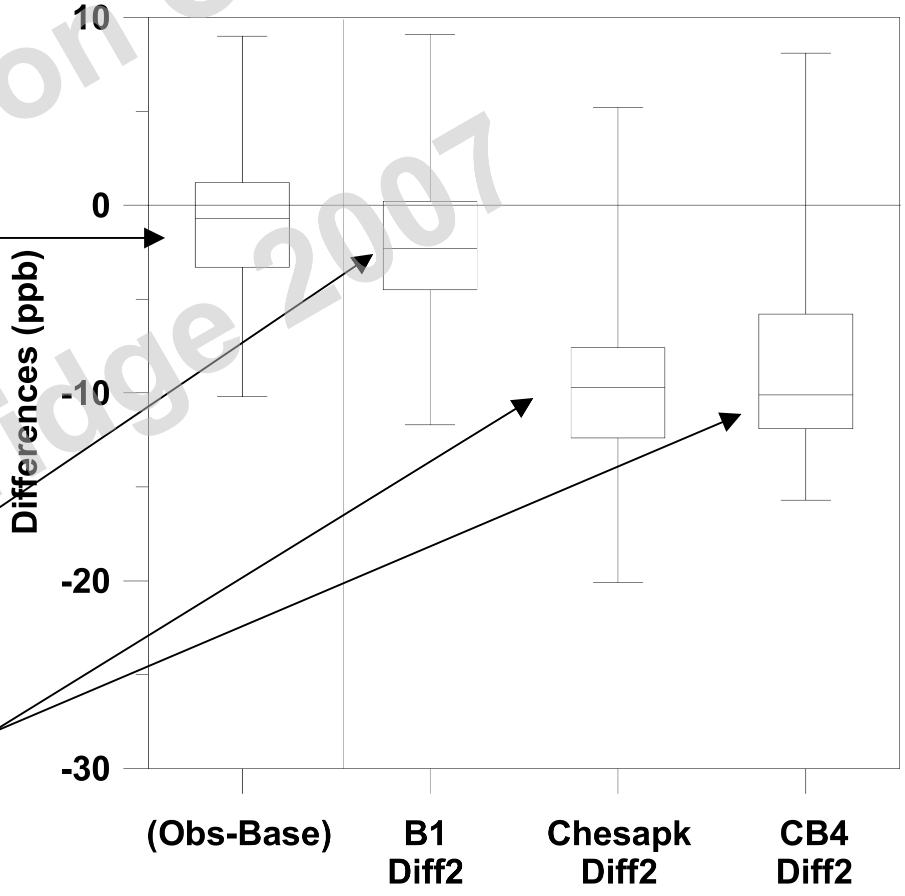
-Compute comparison statistics for each percentile of interest.

> -The best performing model (“Base”) = model with the lowest average for the **abs(Diff-1)**, where **Diff-1 = (obs-model).**
>
> -We check to see if the Base model’s results differ significantly from that observed by inspecting the **distribution of Diff-1 values** to see if the distribution encompasses the value of zero (illustrated on next slide).
>
> -We check to see if the results generated by the other models differ significantly from that generated by the Base model by inspecting the distribution of Diff-2 values to see if the distribution encompasses the value of zero (illustrated on next slide). **Diff-2 = abs(obs-Base) – abs(obs-model)**

# Analyses Conducted At Each Site.



For Site 050350005, for the 90th percentile value, the Base model is determined to be SAPRC (figure to left).

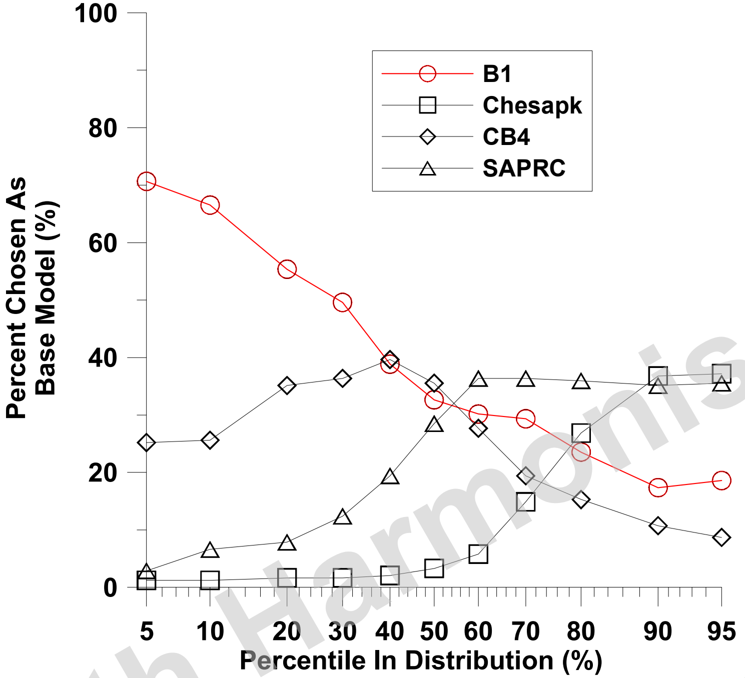


The SAPRC 90th percentile values are found to not differ with the observed 90th percentile values (left-most box plot).

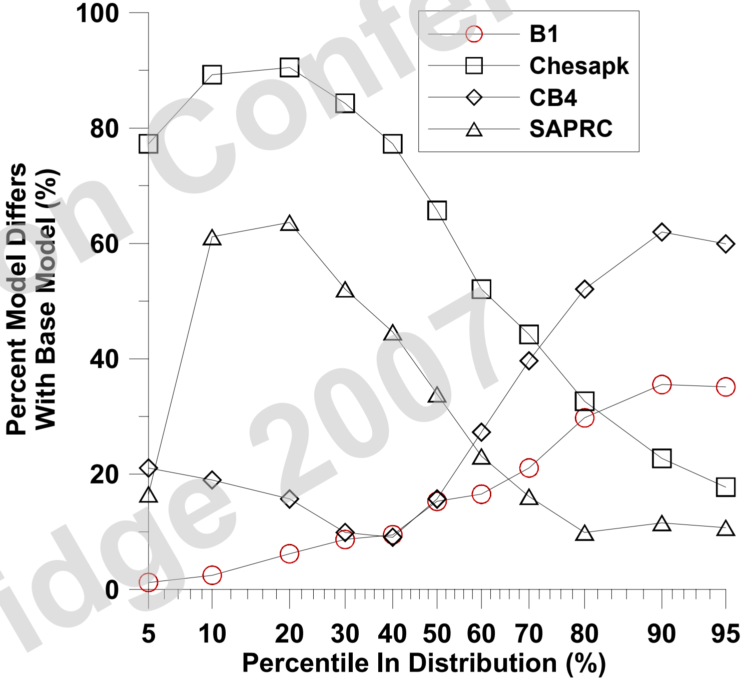
The B1 90th percentile values are found to not differ with the SAPRC 90th percentile values.

The Chesapk and CB4 90th percentile values are found to differ with the SAPRC 90th percentile values.

# Summarizing Over All Sites



How Often A Model Is Selected As the Base Model (Closet to observed)



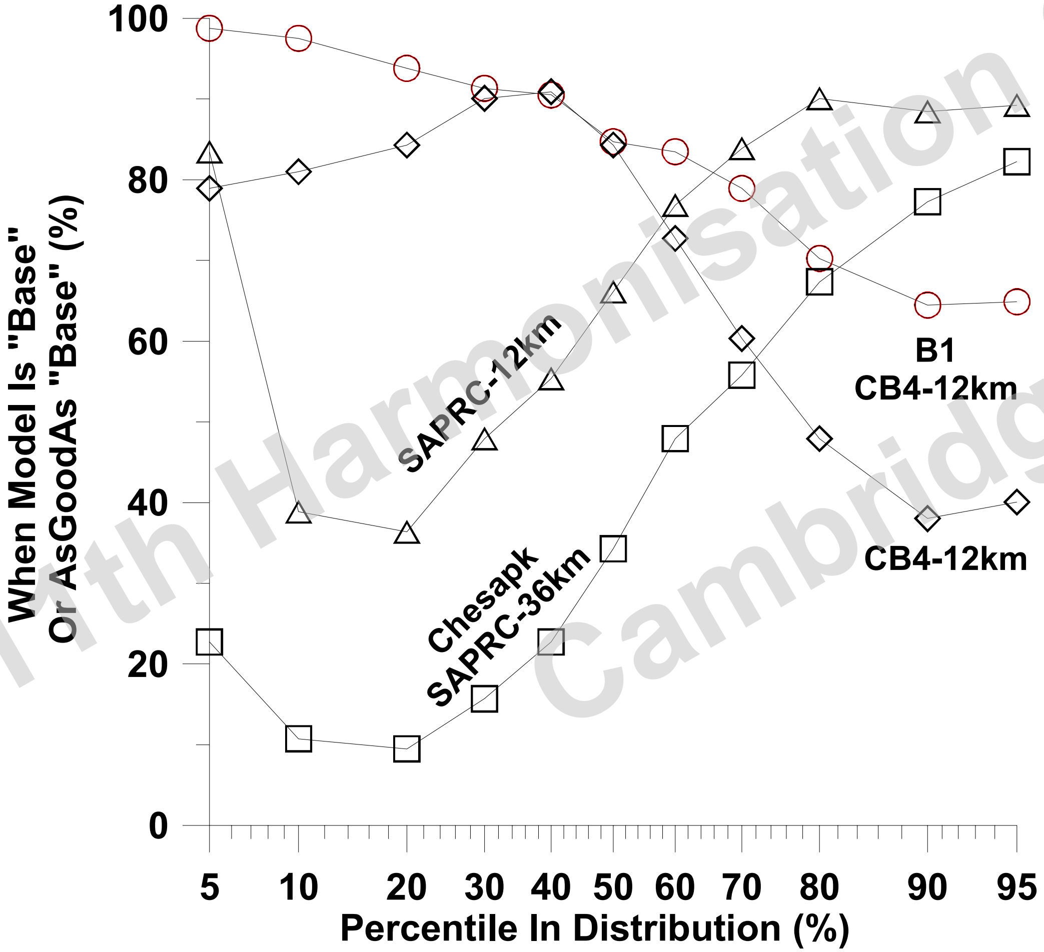
How Often A Model Differs With The Base Model

In a previous slide, the B1 results appeared to be in good correspondence with the observations over the largest range of cumulative frequency distribution.

The left figure confirms that the B1 results are most often selected as the Base model (red line and symbols), except for the higher percentile values.

The right figure confirms that the B1 results do not differ with the Base model's results, except for the highest percentile values.

# Summarizing Further – How Often Is A Model Selected As The "Best" Performing Model Or As-Good-As The Best Performing Model



**Averaging over all percentiles (5% to 95%)**

| **Model** | **Average (%)** |
|---|---|
| B1 | 84 |
| Chesapk | 41 |
| CB4 | 70* |
| SAPRC | 69* |

*Overall score is deceptive!

*What are the causes for the differences seen*? Chemical mechanism is not the total answer, nor is horizontal grid size. Further diagnosis is needed….

# "Further Diagnosis Is Needed"

- How different are the emissions?
  - "Chesapk" used the NEI2001. B1 used the OTC Base.
  - EPA's CB4 and SAPRC used the NEI2002.

- How different are the meteorological inputs?
  - Presumably they all used MM4/FDDA.
  - If there are differences, how and why do these differences occur?
    - Do they explain, in part, the differences seen in the final concentrations simulated?

- How different are the model setups?
  - What are the layer depths, and if differences are seen, what are the consequences?
  - What process characterization options differ, and what are the consequences?

- What analyses are needed to address the questions listed above, and thereby demonstrate the true "cause and effect" relationships?

# Summary

- We have illustrated an operational model evaluation procedure that objectively assesses the relative skill of competing simulations of the variation of the daily maximum 8-hr daily ozone values over a summer season.

- Do we need to further summarize (grand score averaged in some sense over results obtained for 5th-95th percentile values) or are “grand scores” deceptive?

- The coefficient of variation (stdev/avg) of the observed percentile values (determined via bootstrap resampling) is 6-8%.

- Analysis of observed ozone monitored during the summer months of 1975-1976 at St. Louis, MO. Coefficient of variation of point versus 12km averaged 8-hr averages was 12-13%.

- Are significance determinations of practical concern if bootstrap resampling variations are half of that seen in point versus area-averaged 8-hr ozone values (Principle #4)?
  - Is there need to modify the significance tests to account for point versus volume differences?
  - If so, how is this to be done?

# This is the 11-th Harmonization Conference whose purpose is to promote “harmony" in the methods used in air quality assessments

- **My view on how to “harmonize” or bring “into accord or acceptance.”**

**1) Define “Standardized tasks” that are within the scope and capability of the modeled physics and inputs:**

- Since models simulate the “average” to be seen, the “tasks” should not involve extreme values that are dominated by stochastic effects (even if this is the most often use of the model in practice).
- “Skill” is a relative concept whose definition results from an inter-comparison of competing models (e.g. Olympics Decathlon Analogy).

**2) Define “Standardized Tests” to inter-compare models to see if differences seen are “statistically significant”, since models:**

- Only simulate some of the variations to be seen in nature (i.e., the variance simulated will be less than that observed);
- and we must convince ourselves that differences between what is simulated and what is observed are meaningful and not resulting from “unresolved variations” by the model’s physics or inputs.

# References

American Society for Testing and Materials, 2005: Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance (D 6589), (Available at http://www.astm.org), 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428, 17 pages.

Gilliland, A.B., C. Hogrefe, J.L. Godowitch, and S.T. Rao. 2007. Evaluating an Air Quality Model's Ozone Response to Changes in NOx Emissions and Meteorology, to be submitted to *Atmos. Environ.*

Godowitch, J.L.. C. Hogrefe, A.B. Gilliland, and S.T. Rao, 2007. Influence of Point Source NOx Emission Reductions on Modeled Processes Governing Ozone Concentrations and Chemical/Transport Indicators to be submitted to *J. Geophys. Res*.

Nolte, C.G., A.B. Gilliland, and C. Hogrefe, 2007: Linking global to regional models to assess future climate impacts on air quality in the United States: 1. surface ozone concentrations. *J. Geophys. Res.*, in review