Comments on Evaluation Procedures for Air Quality and Meteorological Models

Bob Paine, ENSR

Panel Discussion at EPA's

9th Modeling Conference

October 10, 2008

Outline of Presentation

- Types of evaluation databases
- AERMOD evaluation review
- Evaluation tools
- Cox-Tikvart evaluation procedure
- BOOT/ASTM evaluation procedure (Joe Chang)
- Evaluation databases (Joe Chang)
- Gridded met evaluation

Two Types of Evaluation Databases

- Tracer studies: short-term intensive studies, typically with multiple rows of samplers, each with many sites
  - Can determine plume centerline and plume sigma-y
  - Can determine concentration trend with distance
  - Maximum concentrations on tracer arcs are used for evaluation
  - Can evaluate predictions paired in time and distance
  - Limitation is short duration of study
- Long-term monitoring networks: year-long sampling at a few sites
  - Statistics unpaired in time are necessary; paired in space
  - Limitation is spatial resolution
  - Advantage is large number of hours in database

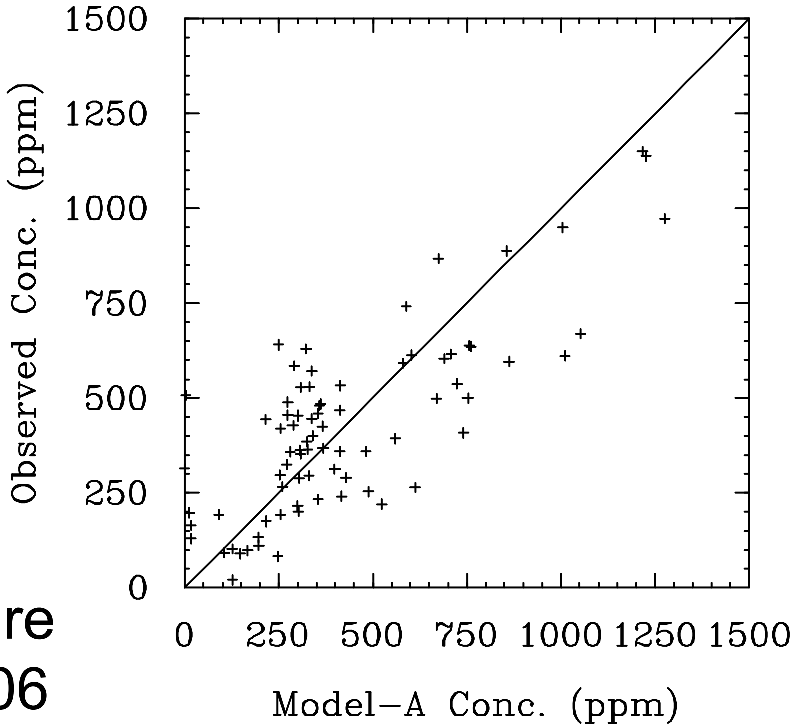
Review of AERMOD Evaluation

- How well does AERMOD predict peak ground-level concentrations used for compliance with AQ standards?
- Is AERMOD's performance significantly better than that of similar models?
- Evaluation databases were a mixture of tracer experiments and long-term studies

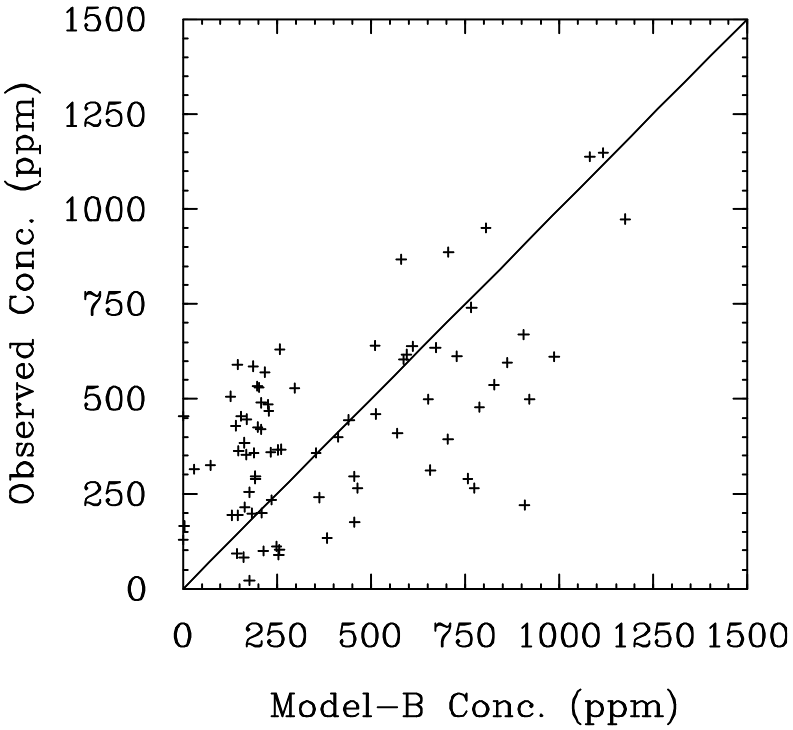
Statistical Evaluation Tools Used for AERMOD

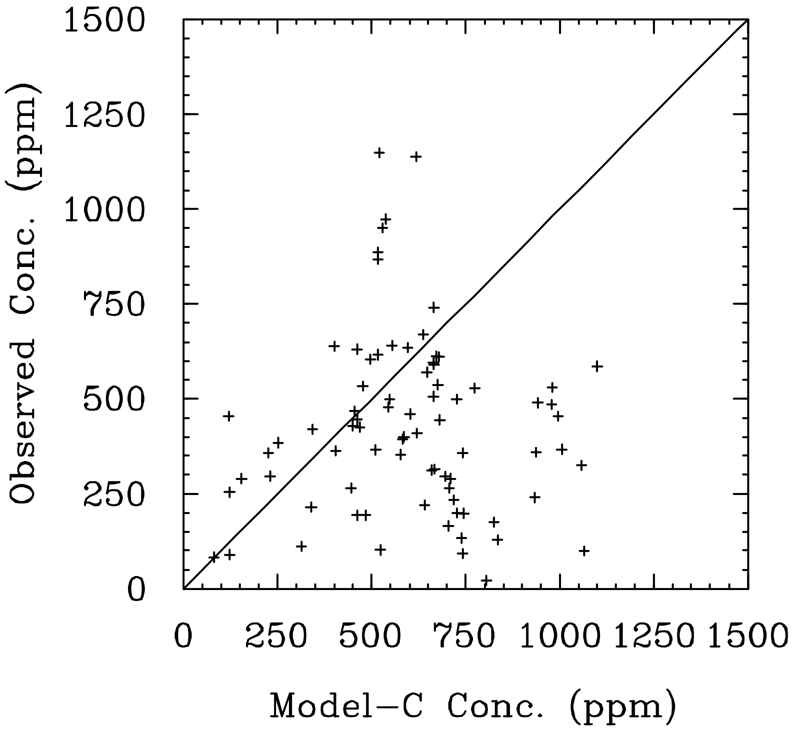
- Plots used extensively; they are often better than “black box” statistics
- Quantile-Quantile (Q-Q) plots: plot pairs of ranked predictions and observations, unpaired in time
  - Can be used for both types of evaluation databases
- Residual plots: plots of ratios of predicted/observed conc vs. downwind distance or wind speed, etc.
  - Generally used only for tracer databases
- Estimates of Robust Highest Concentration, or RHC, that represents a smoothed estimate of the highest concentrations (from Cox-Tikvart evaluation technique)
- Scatterplot (data paired in time and space) – only used for tracer databases

Scatter Plots – Paired in Time and Space



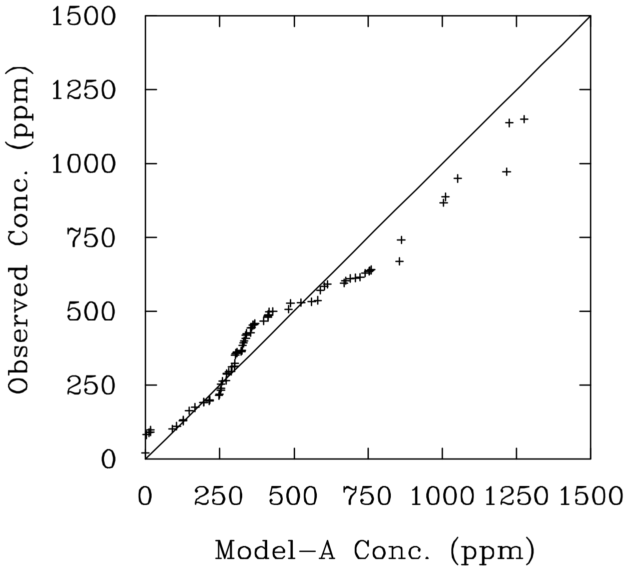
Source: Joe Chang lecture on model evaluation, 2006

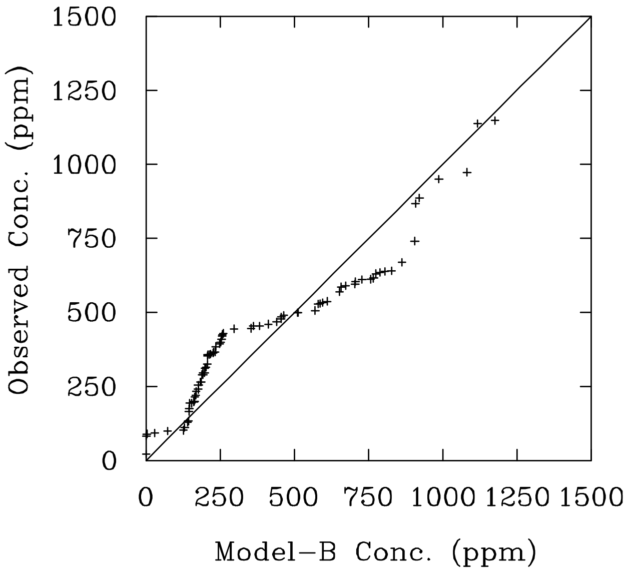


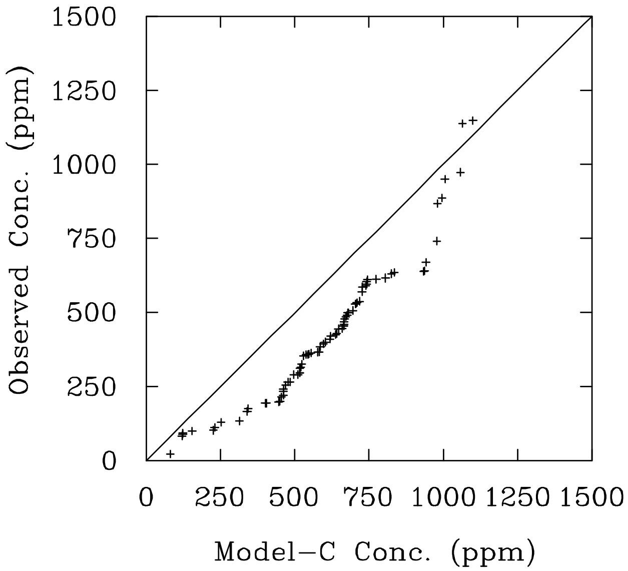


Quantile-Quantile Plot

Source: Joe Chang lecture



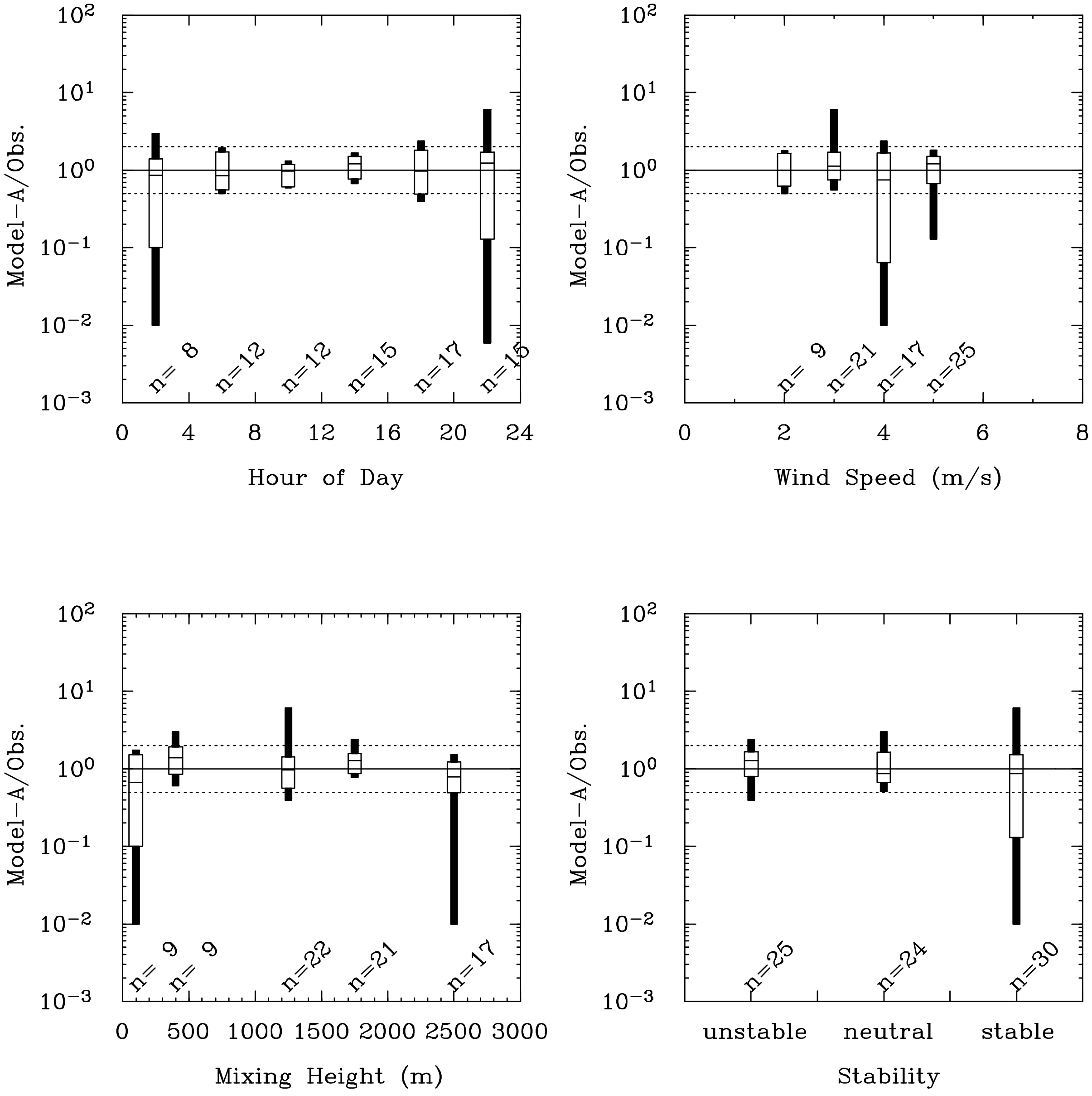




- Observations and predictions are separately ranked
- To see whether CDFs given by observations and predictions are similar
- Does not test ability of model to predict paired in time

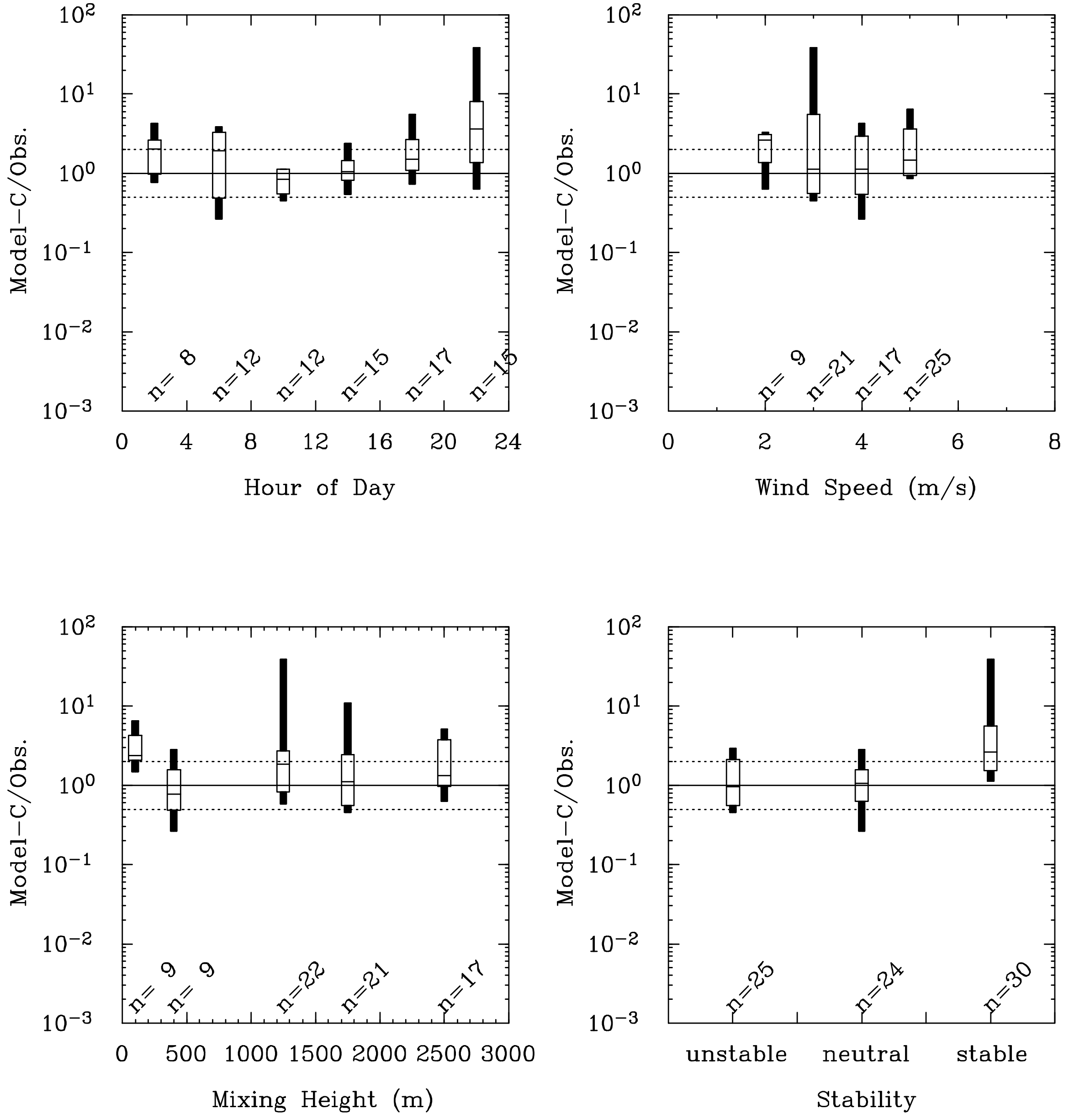
Residual Box Plots for a “Good” Model

Source: Joe Chang lecture



- Plot model residuals, predictions/observations as a function of an independent variable
- Group residuals according to ranges of an independent variable
- Use box plot to indicate the CDF of the n points in each group
- For example, the significant points for each box indicate the 2nd, 16th, 50th, 84th, and 98th percentiles
- A good model should have no trend in model residuals

Residual Box Plots for a “Poor” Model



- A slide trend in model residual is visible

Source: Joe Chang lecture

Important Evaluation Statistic is Fractional Bias

$$F_b = \frac{\overline{C}_0 - \overline{C}_p}{0.5\left(\overline{C}_0 + \overline{C}_p\right)}$$

Co = observed concentration (or Maximum Arcwise Conc. for BOOT/ASTM)

Cp = predicted concentration

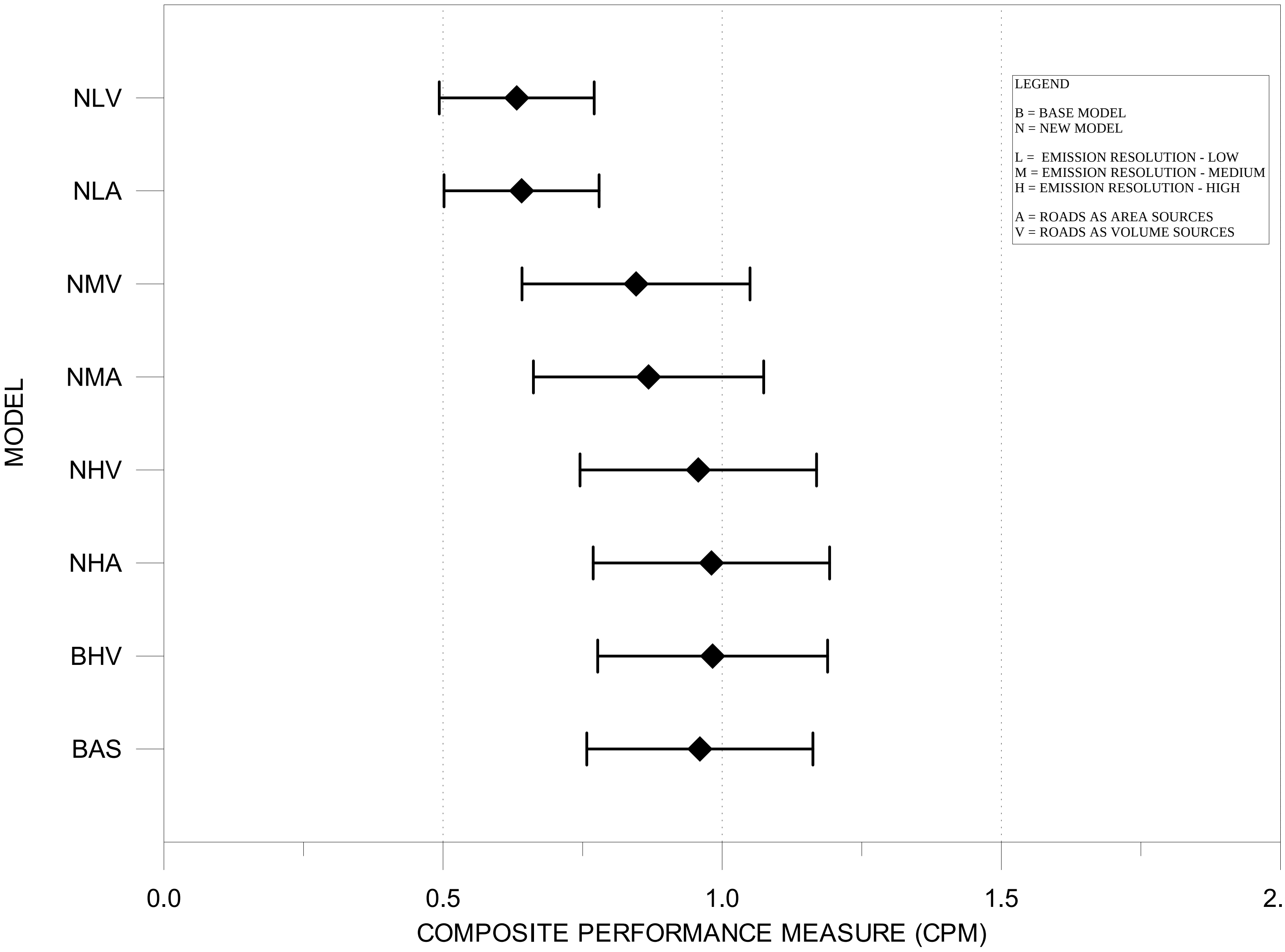
FB of zero is perfect model; +/- 0.67 is within a factor of 2

Major Features of Cox-Tikvart Method

- RHC statistic used
- Resampling of data used to determine confidence interval for differences in performances of models
- Composite performance measure (CPM) combines absolute FBs for several averaging times
- Model Comparison Measure looks at differences in CPM between models to determine statistical significance of differences among models
- Best suited to long-term, sparse network evaluation databases

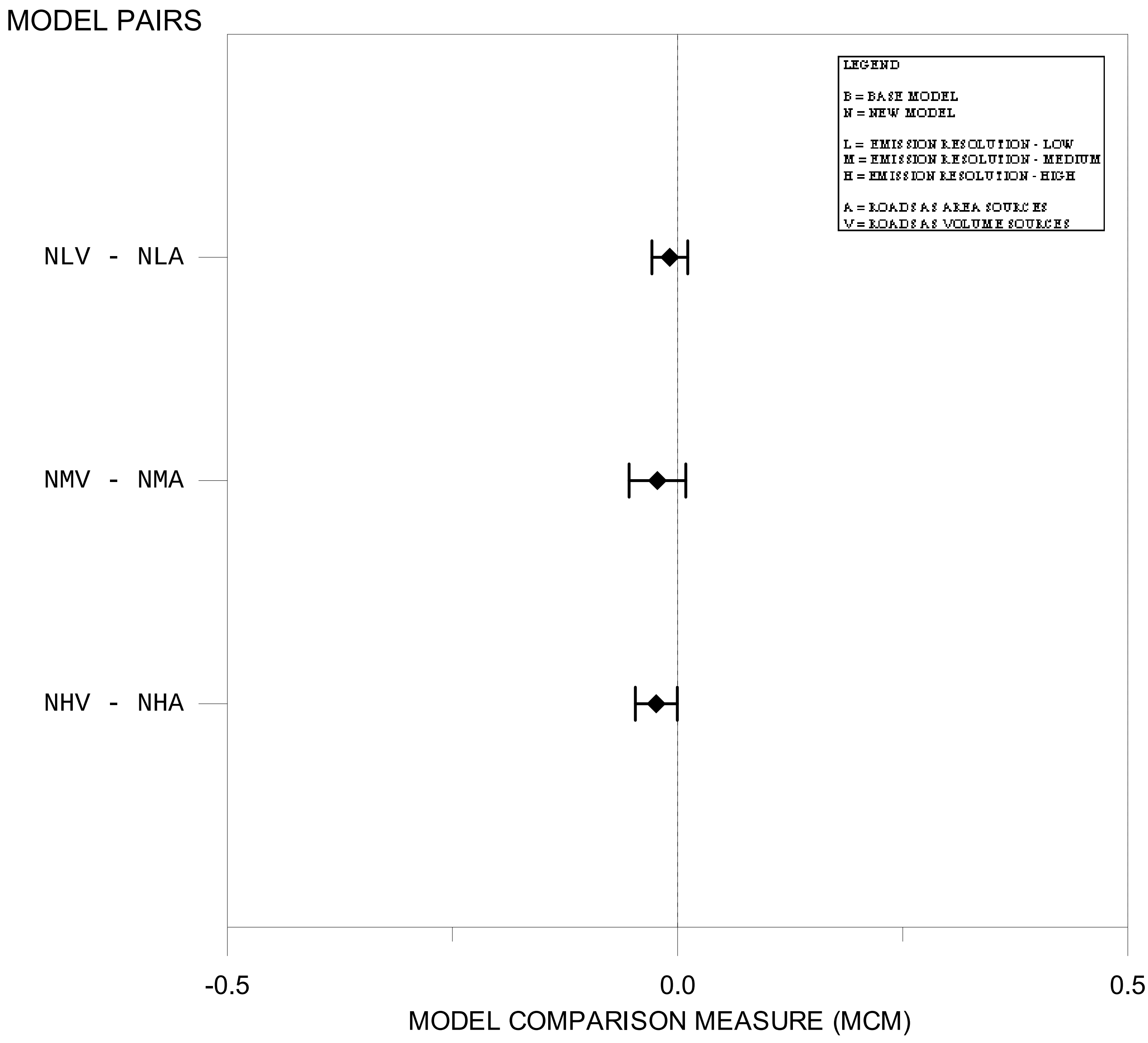
PM-10 Composite Performance Measure (CPM) - ISCST3
With 90% Confidence Limits

(from Brode, 2006)



PM-10 Model Comparison Measure (MCM) - ISCST3
With 90% Confidence Interval

(from Brode, 2006)



BOOT Software Package (slides provided by Joe Chang)

- Developed by Hanna and Chang
- Best suited to tracer databases
- Widely distributed to (> 200) scientists in the field, mainly through the European's *Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes* – Model Validation Kit
- Is generic and can be used to evaluate different kinds of models, different kinds of outputs, and different kinds of data pairings

Primary References

- Chang, J.C., and S.R. Hanna, 2004: Air quality model performance evaluation. *Meteorol. and Atmos. Phys*., **87**, 167-196
- Chang, J. C., 2002: *Methodologies for Evaluating Performance and Assessing Uncertainty of Atmospheric Dispersion Models*. Ph.D. thesis, George Mason University, Fairfax, VA 22030-4444, 277 pp
- These two references lead to numerous other citations

Performance Measures in BOOT

$$FB = \frac{\left(\overline{C_o} - \overline{C_p}\right)}{0.5 \left(\overline{C_o} + \overline{C_p}\right)}$$

Fractional Bias

$$NMSE = \frac{\overline{\left(C_o - C_p\right)^2}}{\overline{C_o}\ \overline{C_p}}$$

Normalized Mean Square Error

$$MG = \exp\left(\overline{\ln C_o} - \overline{\ln C_p}\right)$$

Geometric Mean Bias

$$VG = \exp\left[\overline{\left(\ln C_o - \ln C_p\right)^2}\right]$$

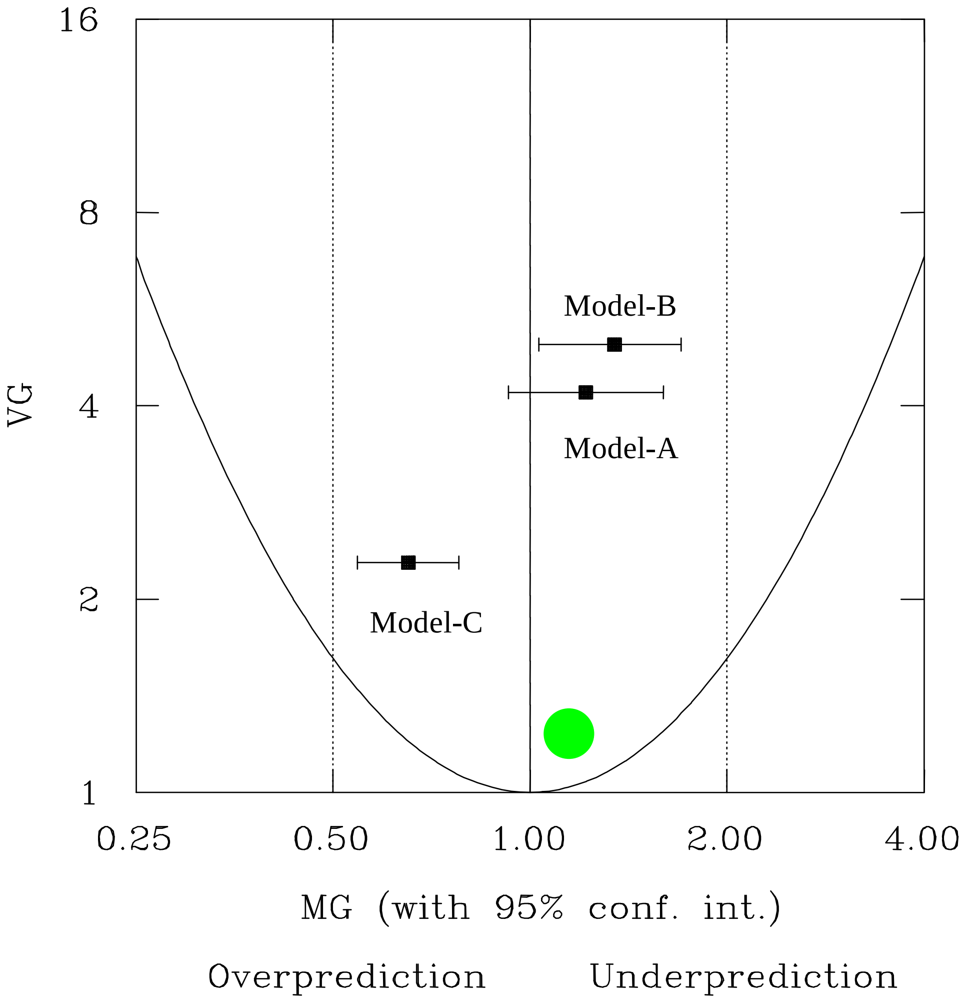
Geometric Variance

$$FAC2 = \text{\% of data that satisfy } 0.5 \le \frac{C_p}{C_o} \le 2.0$$

$$R = \frac{\overline{\left(C_o - \overline{C_o}\right)\left(C_p - \overline{C_p}\right)}}{\sigma_{C_p}\sigma_{C_o}}$$

Correlation Coefficient

Examples of BOOT Performance Plot



- A nice way to plot MG/VG (or FB/NMSE) at the same time
- A perfect model is located at the center of the x-axis (green dot)
- MG for Model-B and Model-C are significantly different from 1.0

What Are “Observations”?

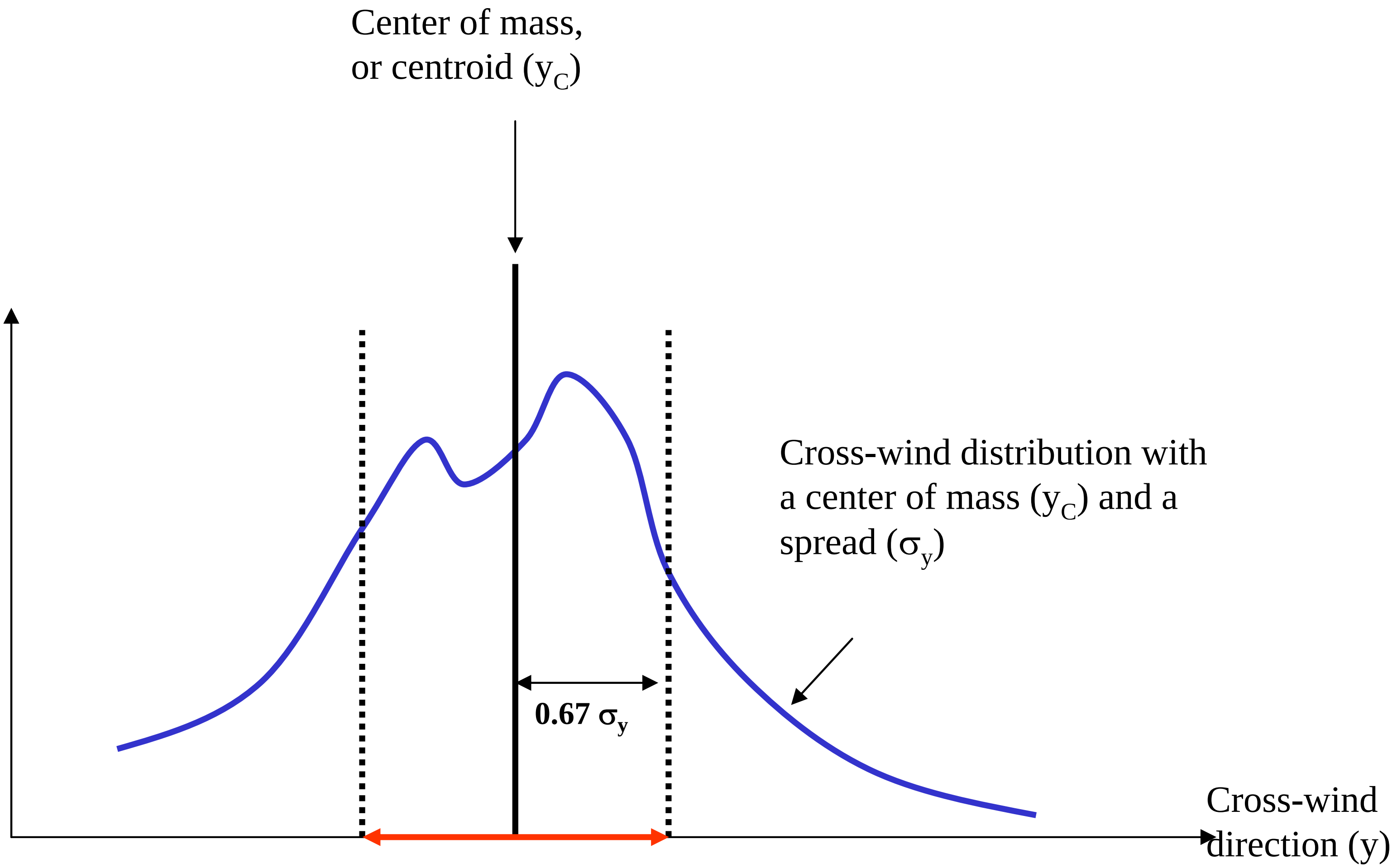
- Observations can be
  - Directly measured by instruments
  - Products of other models or analysis procedures
- Direct observations are *snapshots* of an ensemble, but model predictions often represent ensemble averages

ASTM (American Society for Testing and Materials) Procedure – Similar to BOOT

- Observations are snapshots (ensemble realizations)
- Model predictions are ensemble averages
- The two cannot be directly compared
- In order to compare model predictions to observations, some sort of averaging must first be performed
- ASTM suggests that this averaging be done over *regimes* of similar conditions (*e.g.*, for downwind distance or atmospheric stability)

ASTM Procedure

- For short-range dispersion experiments where samplers are arranged in arcs, ASTM procedure also suggests near-centerline concentrations as representative of centerline values – creates Gaussian fit

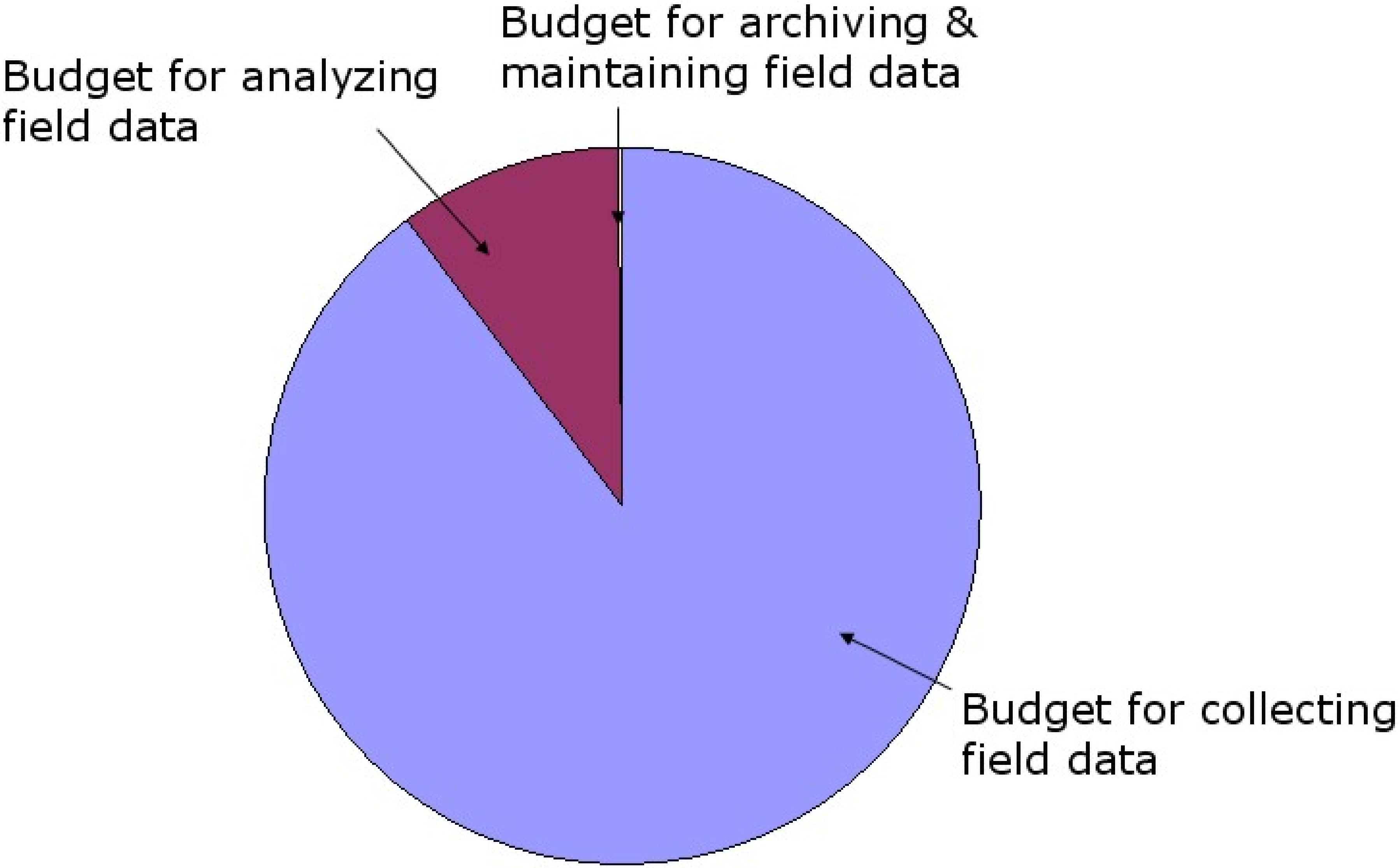


Center of mass, or centroid (y_C)
Cross-wind distribution with a center of mass (y_C) and a spread (σ_y)
0.67 σ_y
Cross-wind direction (y)

Issues With ASTM Procedure

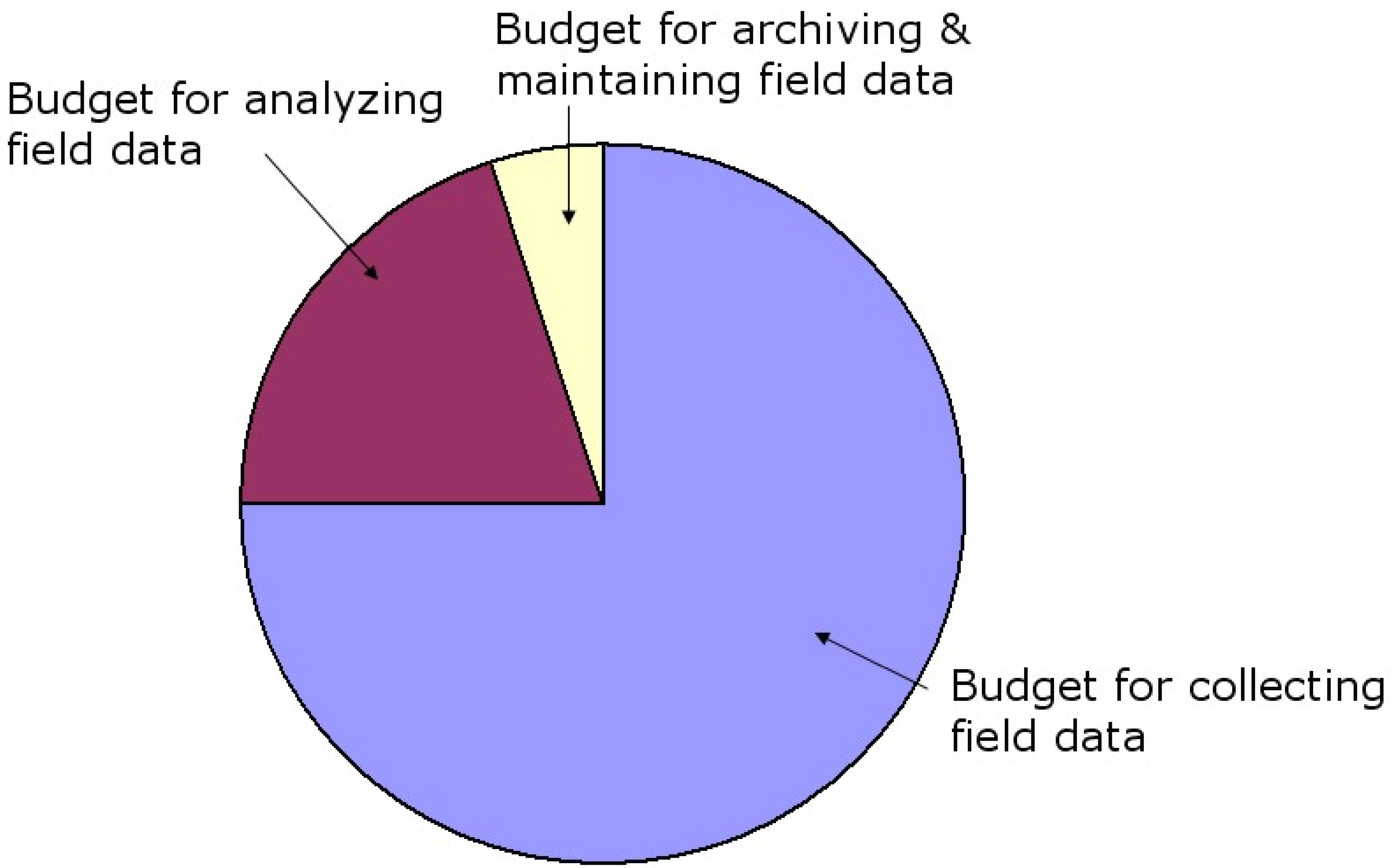
- Results sensitive to how the limited regimes are defined
- Has so far only been demonstrated for short-range dispersion experiments with concentric sampling arcs
- Not clear how the procedure should be applied to the evaluation of 3-D Eulerian air quality models, where predicted concentrations represent averages over a grid volume, but observed concentrations represent point measurements

The Reality



Budget for analyzing field data
Budget for archiving & maintaining field data
Budget for collecting field data

A Better Scenario



See > 100 database references from Joe Chang at
http://www.ofcm.gov/homeland/gmu2005/Presentations/09-Chang%202005%20GMU-OFCM%20Panel.ppt

Evaluation of Gridded Meteorological Data

- Gridded met data should not be used until thoroughly evaluated with independent data
- There may be situations with poor met performance (e.g., complex terrain)
- Conditions of concern for dispersion modeling:
  - Low wind frequency
  - Underestimation of wind speeds aloft (e.g., low-level jet)
  - Wind rose misrepresentation
- Sources of data for testing
  - Need to find tall tower data, not just surface data
  - Private industrial met towers
  - Numerous wind energy assessment towers