Performance Evaluations Metrics for Long Range Transport Models

Evaluation Paradigm for Long Range Transport Models

- LRT models play a unique role in air quality modeling. This class of models plays several roles.
  - Emergency response modeling
  - Class I increments and visibility
- Requires additional level of skill to reflect time and space considerations of LRT model use
  - Statistical measures should examine spatiotemporal pairing ability of LRT models.

Evaluation Goal

- Develop meteorological and tracer databases for evaluation of long range transport models.
- Develop a consistent and objective method for evaluating long range transport (LRT) models used by the EPA.
- Promote the best scientific application of models based upon lessons learned from evaluations and reflect this in EPA modeling guidance.

Irwin Methodology

Spatial	Temporal	Performance
Azimuth of Plume Centerline	Plume Arrival Time	Crosswind Integrated Concentration
Plume Sigma-y	Transit Time on Arc	Observed Maximum

Method:

- Compute *n*-hour average for each receptor on arc (ArcAverage program)
- Use trapezoidal integration to ‘fit’ average plume on arc (PLMFIT program)

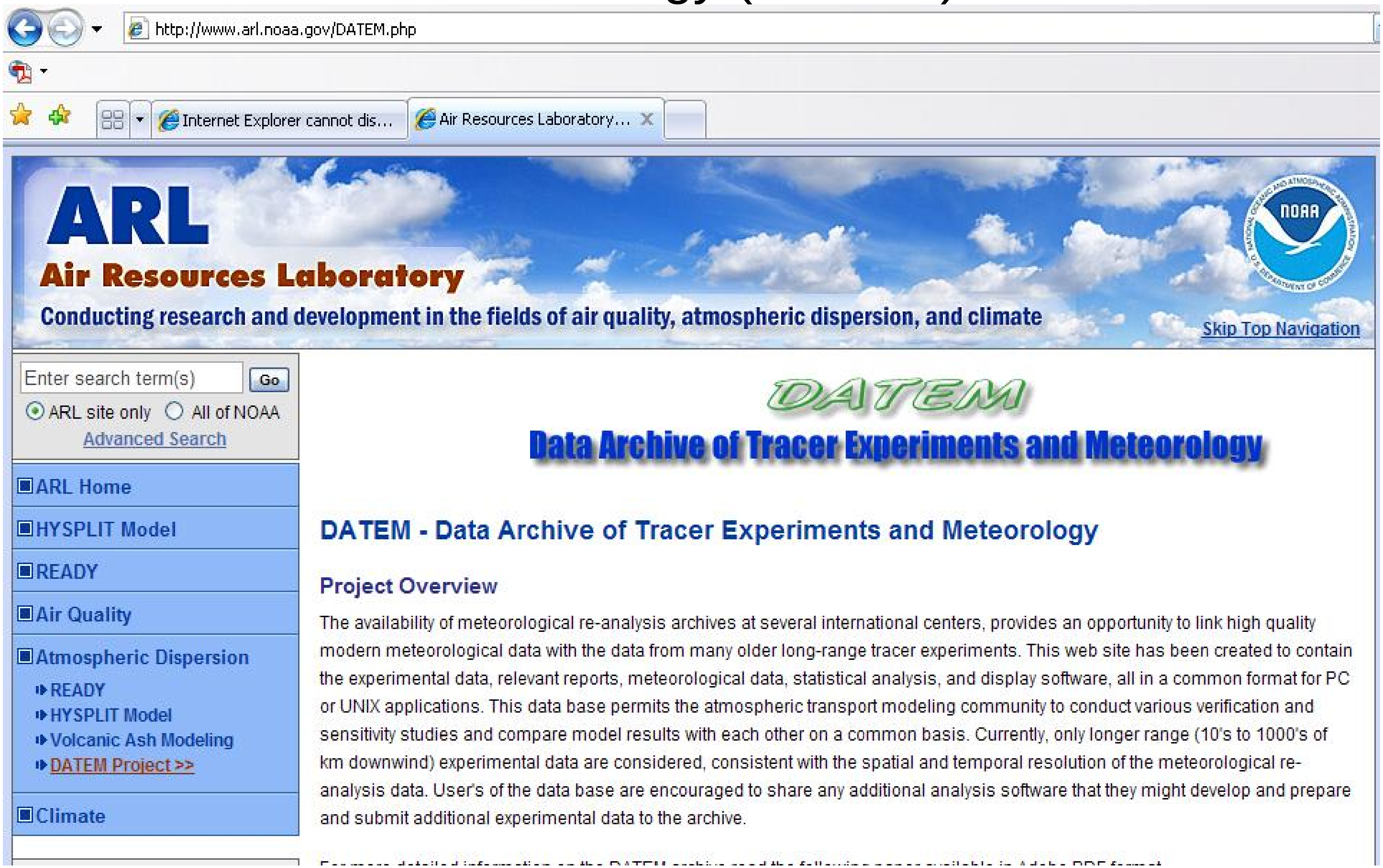
Background from Original Performance Evaluations

- Measures employed by Irwin (1998) and EPA (1998) provide useful diagnostic information about the performance of LRT modeling systems such as CALPUFF, but they do not always lend themselves easily to spatiotemporal analysis or direct model intercomparison.
- For tracer studies such as the Great Plains Tracer Experiment, Savannah River, and INEL74 where distinct arcs of monitors were present, the Irwin evaluation approach was used.
- In addition to the Irwin methodology, EPA augmented statistical measures focusing upon spatiotemporal comparisons of model-observation pairings.

Statistical Evaluation Methodology

- The model evaluation methodology employed for this project was designed following the procedures of Mosca et al. (1998) and Draxler et al. (2001).
- Statistical measures fall into four broad categories
  - Scatter
  - Bias
  - Spatial
  - Cumulative distribution
- NOAA ARL DATEM performance evaluation program (STATMAIN) augmented by EPA with additional spatial statistics for false alarm rates, probability of detection, and threat score.

NOAA's Data Archive of Tracer Experiments and Meteorology (DATEM)



Statistical Measures - Scatter

$$FOEX = \left[\frac{N_{(P_i > Ni_i)}}{N} - 0.5 \right] \times 100\%$$

- Factor of Exceedance (FOEX)

$$FA\alpha = \left[\frac{N\left(y - y0 = [x - x0]\alpha\right)}{N} \right] \times 100$$

- Factor of 2/5 (FA2/5)

$$NMSE = \frac{1}{N\overline{PM}} \sum \left(P_i - M_i\right)^2$$

- Normalized Mean Square Error (NMSE)

$$R = \frac{\sum_i \left(M_i - \overline{M}\right) \bullet \left(P_i - \overline{P}\right)}{\left[\sqrt{\sum \left(M_i - \overline{M}\right)^2}\right]\left[\sqrt{\sum \left(P_i - \overline{P}\right)^2}\right]}$$

- Pearson's Correlation Coefficient (R)

Statistical Measures – Cumulative Distribution

- Kolmogorov-Smirnov Parameter $KS = MAX\left|C\left(M_k\right) - C\left(P_k\right)\right|$

Statistical Measures - Bias

$$\overline{B} = \frac{1}{N}\sum_i (P_i - M_i)$$

- Mean Bias (B)

$$FB = 2\overline{B} / \left(\overline{P} + \overline{M}\right)$$

- Fractional Bias (FB)

Spatial Statistics

$$FMS = 100\,\frac{N(p>0)\cap N(m>0)}{N(p>0)\cup N(m>0)}$$

$$FAR = \left(\frac{a}{a+b}\right)\times 100\%$$

$$POD = \left(\frac{b}{b+d}\right)\times 100\%$$

$$TS = \left(\frac{b}{a+b+d}\right)\times 100\%$$

- Figure of merit in Space (FMS)
- Additional EPA Metrics
  - False Alarm Rate (FAR)
  - Probability of Detection (POD)
  - Threat Score (TS)

a=forecast, not observed

b=forecast and observed

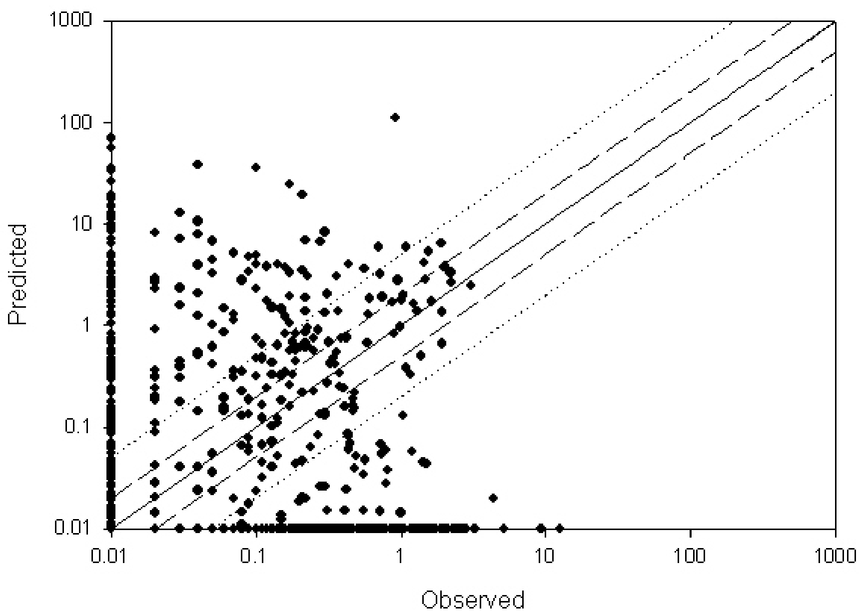
d=observed, not forecast

Statistical Measure – Final Model Rank

$$RANK = R^2 + \left(1 - \left|FB/2\right|\right) + FMS/100 + \left(1 - KS/100\right)$$

- Model Rank (RANK) – measure of model "success" across each of the four broader statistical categories.
- Range from 0 to 4, with 0 poorest and 4 best performance.
- Unique measure allows for direct intercomparison amongst models and summarizes success of model into a single number easily relatable

Example Performance Evaluation – FLEXPART 6.2



-
- EXP_DATA results file: gbl1_001.txt
- Model variation: 001 Tracer number: 1 Station select: All
- -----------------------------------
-
- 3105 Unaveraged data points for processing
- 0.00 Percentile input for zero measured
- 0.00 Zero measured concentration value
-
- 0.00 Correlation coefficient
- 0.06 T-value (|Slope|/Standard Error)
- 269.56 Average measured concentration
- 874.65 Average calculated concentration
- 3.24 Ratio of calculated/measured
- 118.44 Normalized mean square error
- 1104 Number of pairs analyzed
-
- 605.09 Average bias [(C-M)/N]
- 197.29 Lo 99 % confidence interval
- 1012.90 Hi 99 % confidence interval
- 1.06 Fractional bias [2B/(C+M)]
-
- 30.07 Fig of merit in space (%)
-
- 0.54 False Alarm Ratio
- 0.27 Probability of Detection
- 0.20 Threat score at 0.1 ngm-3 threshold
-
- -18.30 Factor exceeding [N(C>M)/N-0.5]
- 5.53 Percent C/M ± 2
- 12.77 Percent C/M ± 5
- 30.07 Percent M>0 and C>0
- 55.16 Percent M>0 and C=0
- 14.76 Percent M=0 and C>0
-
-
- 51.00 Kolmogorov-Smirnov Parameter
-
- 1.26 Final rank (C,FB,FMS,KSP)