Consolidating tools for model evaluation

Helge Rørdam Olesen
National Environmental Research Institute (NERI)
Denmark

Joseph C. Chang
George Mason University, Virginia
USA

What is the status of tools for model evaluation?

The presentation focuses on the
Model Validation Kit

Also provides some information on the
ASTM methodology
for statistical evaluation of dispersion models.

Reason for giving the presentation

- There may easily be confusion over which tools are available.
- A new version of the Model Validation Kit is now available.

Features of the Model Validation Kit

- Addresses classic single-source problem.
- Four field data sets.
- BOOT software for statistical performance evaluation.
- SIGPLOT software. Option for exploratory data analysis.
- Utilities to facilitate use of the software; define standard set of output plots etc.
- The *Dispersion Visualisation Tool* – utility to inspect Kincaid tracer data.
- Video film from Kincaid.

History

- Introduced at Manno workshop in 1993.
- Official version from Mol workshop in 1994.
- Supplement added in 1997.

- 250 hard copies distributed since 1993.

- Version 2.0 released in October 2005.

What is new in version Version 2.0?

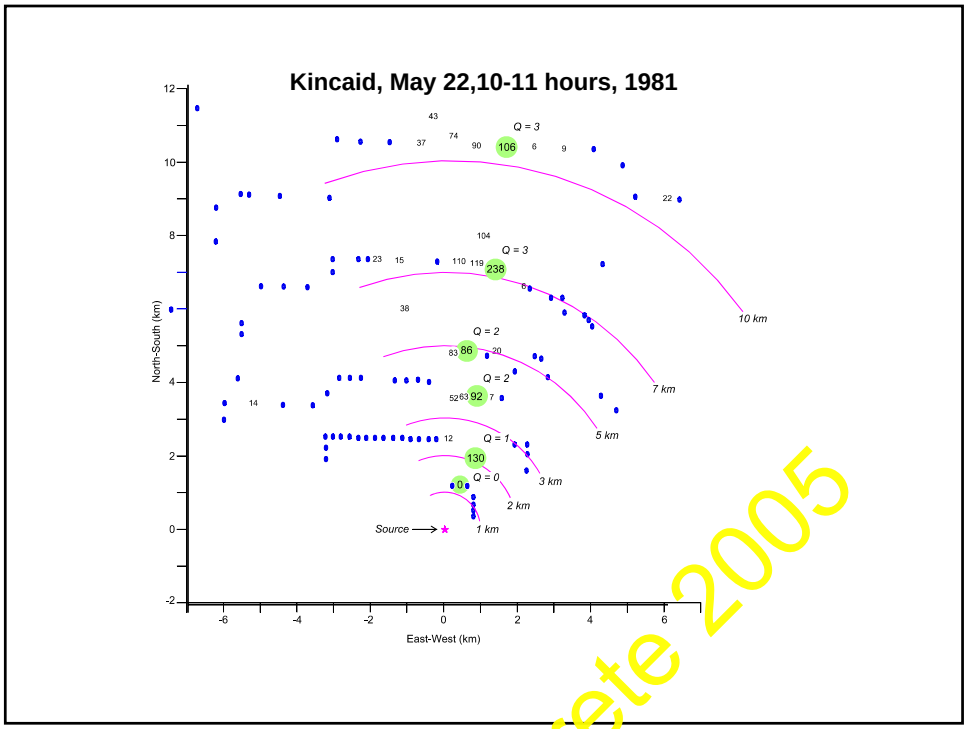
- Some elements of the old version didn't work in a modern Windows environment
- The documentation is considerably improved.
- Available on the web
- New version of BOOT included.
- Additional features included:
  - The Dispersion Visualisation Tool
  - Video film from Kincaid
- More info added.

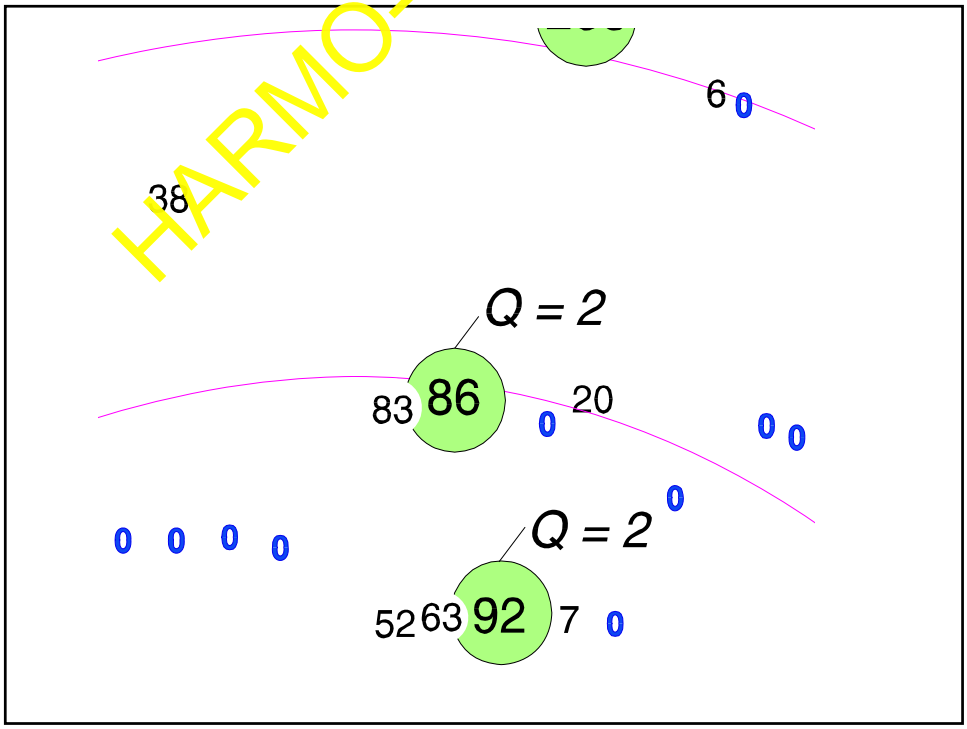
Data sets (1)

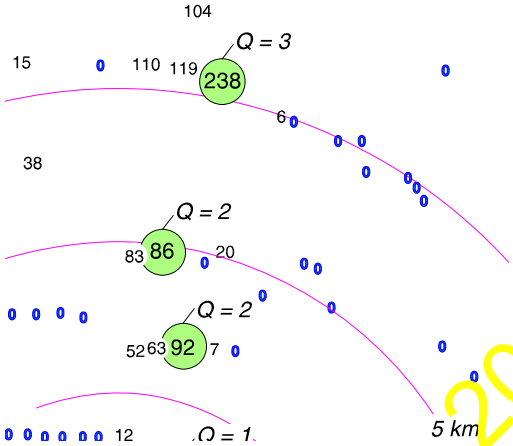
- **Each data set involves a single source in homogeneous terrain.**
- **Concentration variables considered:**
  a) arc-wise maxima
  ***b) cross-wind integrated concentrations*.**

Data sets (2)

- **Kincaid. Buoyant, 189 m source. 170 experiments. Arc-wise maxima.**
- **Copenhagen. Passive, 115 m source. 9 experiments. Crosswind int. + arc-wise maxima.**
- **Lillestrøm. Passive, 36 m source. 8 experiments. Crosswind int. (+ arc-wise maxima).**
- **Indianapolis. Urban, buoyant, 84 m source. 171 experiments. Arc-wise maxima.**







Useful feature: a quality index

- Used for arc-wise maximum concentrations in Kincaid and Indianapolis
- The quality index has values of 0, 1, 2 and 3, with 2 and 3 representing the most reliable data. Comparison studies of observed data with model results should in general be conducted with a quality indicator of 2 or 3.
- Subsets of data can be selected in a well-defined manner.

Technical Descriptions and User's Guide for the BOOT Statistical Model Evaluation Software Package, Version 2.0

by

Joseph C. Chang[1,2] and Steven R. Hanna[3]

[1]Comprehensive Atmospheric Modeling Program
School of Computational Sciences
George Mason University
4400 University Drive, MS 5B2
Fairfax, VA 22030-4444

[2]also affiliated with
Homeland Security Institute
2900 South Quincy Street, Suite 800
Arlington, VA 22206-2231

[3]Harvard School of Public Health
Landmark Center, Room 404J
401 Park Drive
Boston, MA 02215-0013

July 10, 2005

Statistical performance measures (1)

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

Fractional bias, measuring systematic error on a linear scale

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

Geometric mean bias, measuring systematic error on a log scale

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

Normalized mean square error, measuring systematic and random error on a linear scale, heavily biased by large values

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

Geometric variance, measuring systematic and random error on a log scale, heavily biased by small values

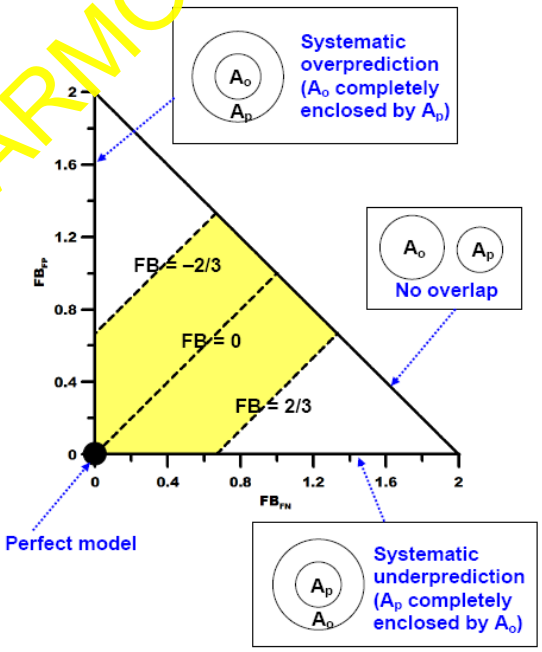
$$R = \frac{\overline{\left(C_o - \overline{C_o}\right)\left(C_p - \overline{C_p}\right)}}{\sigma_{C_p}\sigma_{C_o}}$$

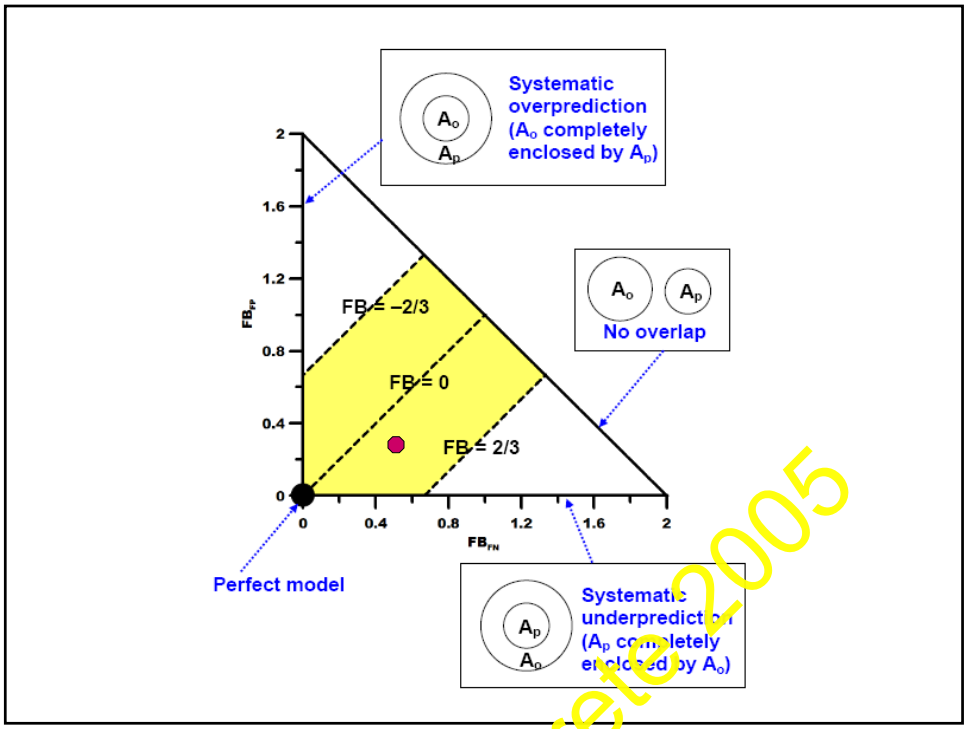
Correlation coefficient, not very robust

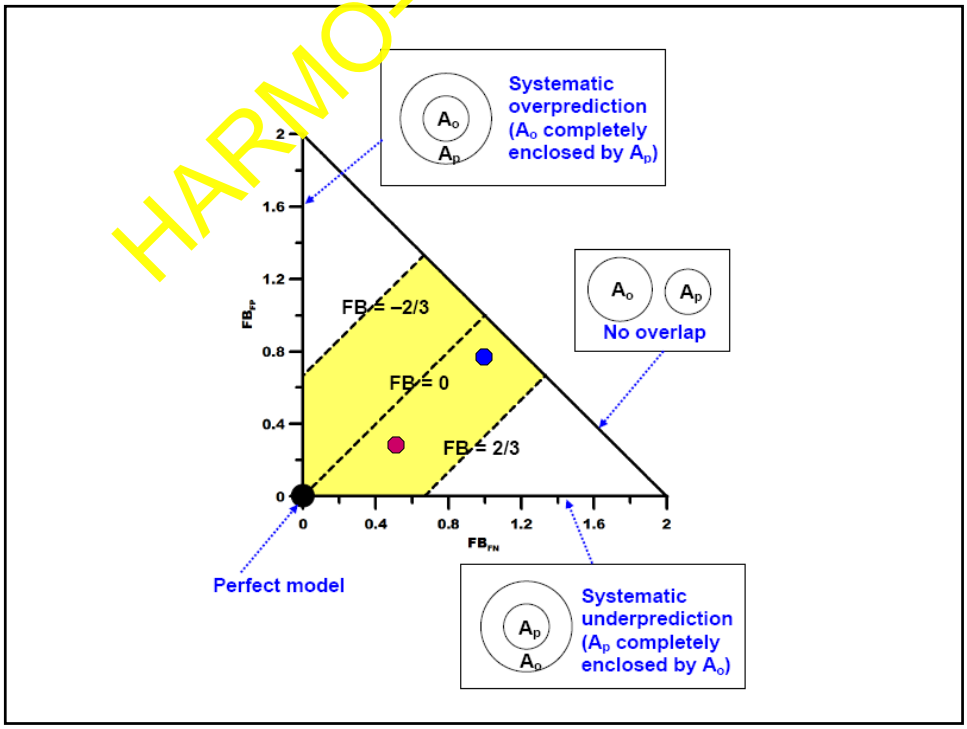
$$FAC2 = \text{fraction of data that satisfy } 0.5 \le \frac{C_p}{C_o} \le 2.0$$

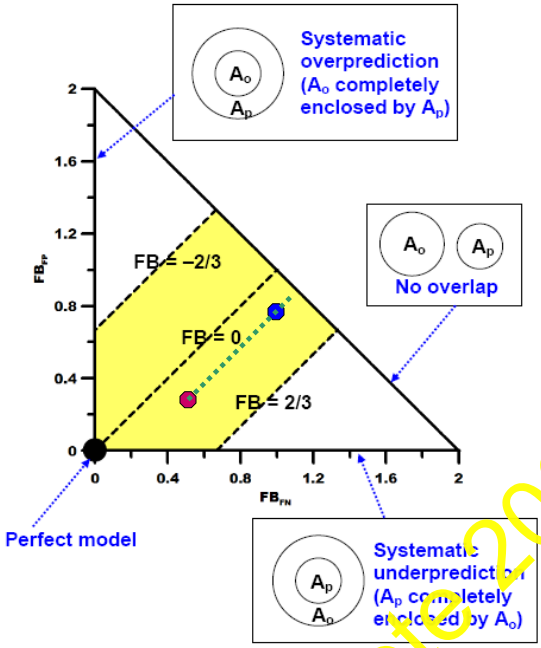
Statistical performance measures (2)

- **FB (fractional bias): Of limited value because overpredictions and underpredictions compensate each other.**
- **A useful extension:**
  - **FB_{FN}** (false negative) considers only underpredictions
  - **FB_{FP}** (false positive) considers only overpredictions









Statistical performance measures (3)

- Two-dimensional Measure of Effectiveness (MOE) closely related to FB_{FN} and $FB_{FP.}$
- Absolute fractional bias (AFB) is the sum of FB_{FN} and $FB_{FP.}$
- BOOT computes:
- FB_{FN}, FB_{FP}, MG_{FN}, MG_{FP}, MOE_{FN} and $MOE_{FP.}$

BOOT features (continued)

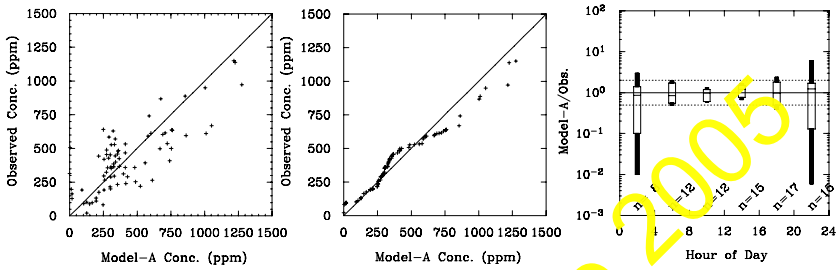
- **The statistics part of the ASTM methodology is implemented.**
  However, the preparatory work of regime definitions and data stratification is <u>not</u> part of BOOT.

Important

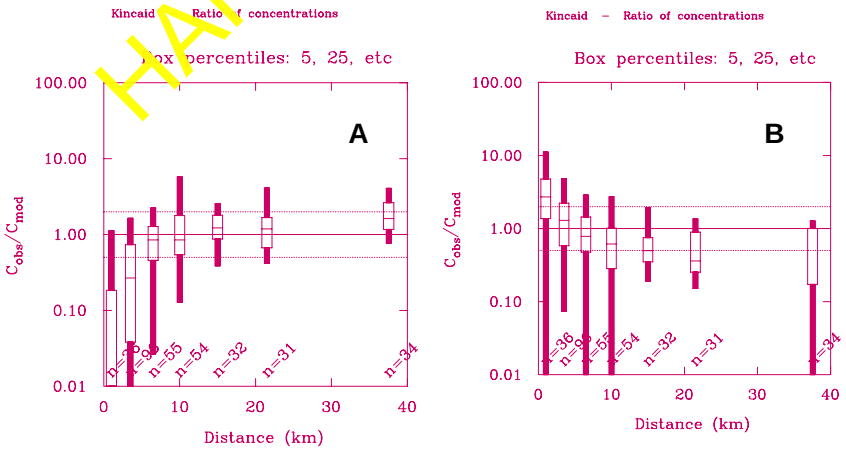
- **Statistical performance measures provide only limited information.**
- **Need for exploratory data analysis!**

Exploratory data analysis

- **Scatter plot (left)**
- **Quantile-quantile plot (middle)**
- **Residual plot (right)**



Behaviour of models can easily be compared



SIGPLOT software

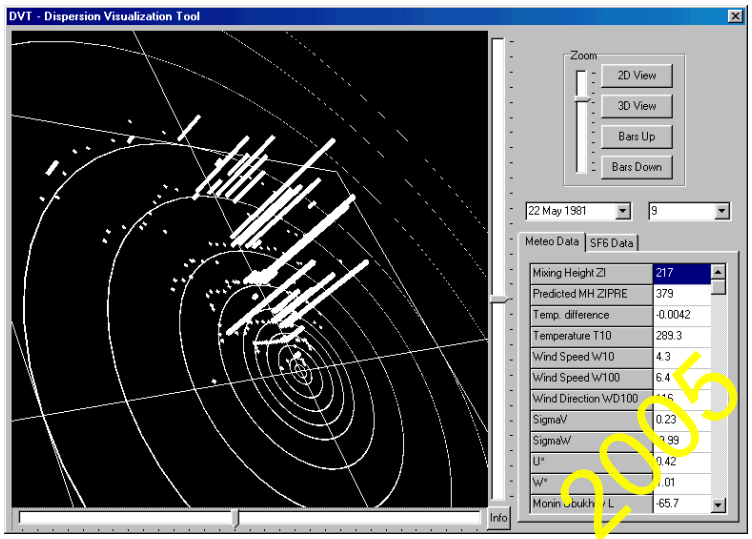
- Offered as an option.
- Pretty old – requires simulation of DOS environment
- but it works, and produces plots that are easy to compare with those others have produced.
- Produces specialised plots, such as box plots for exploratory analyses.
- The Model Validation Kit contains the necessary template files and tools, and it provides a step-by-step explanation of the approach.

Dispersion Visualisation Tool

Created by Alexandar Markoski, University of Bitola, FYROM

Accompanied by measurements from the Kincaid experiment (1980)

Dispersion Visualisation Tool

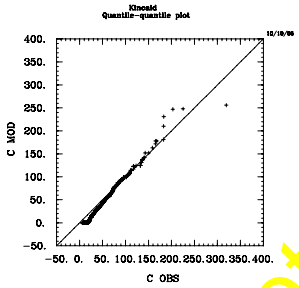


Video film from Kincaid

- 8 video clips in mpg format, of duration 0-2 minutes each
- Most of them are time-lapse sequences

Model Validation Kit - limitation

- A notable limitation of the standard procedure prescribed in the kit:
- It does not explicitly address the stochastic nature of atmospheric dispersion.
- Note that quantile-quantile plots should *not* be expected to give a one-to-one correspondance.



Model Validation Kit - the plus side

- It is straightforward to apply.
- Results are produced in a standardised way.
- Residual plots are useful.

- but all results should be interpreted with care.

An alternative approach adopted in the ASTM *Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance*

- **The new BOOT software allows this approach. However, the Model Validation Kit does not contain ready-to-use utilities to prepare observed data for this purpose.**
- **An an alternative package exists. Prepared by John Irwin.**

Fundamental premise of ASTM approach

- **Observations and predictions should *not* be compared directly.**
- **Instead, the comparison takes place within *regimes*.**
- **Regimes can, e.g., be defined according to distance to the source and atmospheric stability.**
- **Performance measures are calculated based on *regime averages* - rather than values for individual experiments.**

ASTM package (Irwin)

- **Software**
- **Documentation**
- **Three data sets:**
  - **Prairie Grass**
  - **Kincaid**
  - **Indianapolis**

- **Focus is on Near Centerline Concentrations.**
  Data are not quality flagged, but the software performs certain automatic checks.

Issues deserving attention (1)

- **Regimes can be defined in many different ways. If some very different scenarios are grouped together in the same regime, results may be misleading.**

- **The procedure considers near-centreline concentrations. In the current implementation it is problematic that near-centreline concentrations are compared to a model prediction in the *exact* centerline.**
  By definition a centerline concentration is higher than near-centerline values.

Issues deserving attention (2)

- **The basic assumption that model results should fit observations may not always be warranted. It is vital to assure proper quality of observed data.**
- **Problems with the observed data or the way they are interpreted may easily pass unnoticed if you just feed experimental data into a statistical "blackbox".**
- **Use of a quality indicator could alleviate such problems .**

Web addresses:

- **Model Validation Kit:**
  www.harmo.org/kit
- **John Irwin's package implementing the ASTM methodology:**
  www.harmo.org/astm

In conclusion

- None of the evaluation protocols – neither the one used in the Model Validation Kit nor the one used in the ASTM approach – are so robust that they can be applied without reservation.
  Often, they will lead to 'inconclusive conclusions'.
- Nevertheless, model evaluation based on the existing tools is extremely useful to promote the quality of models. Many model weaknesses can be revealed.
- There is still a lack of data sets that have been quality checked and carefully prepared for model evaluation.

Processing of input data is far from trivial !

Examples:

- How should arc-wise maxima be determined?
- How about near-centerline concentrations?
- How about cross-wind integrated conc.?

Some experiences:

- Take care!
- Identify pitfalls!
- Use quality indicators to define good-quality subsets of data

A lot of work ahead (1)

- The moral of the story is that as a producer of data you have to work your way through the data and test things out; you should not just take a data set from the shelf and distribute it, assuming that your job is over.
  When working through the data, you will encounter numerous problems on your way, both tiny problems and larger.
  All of these problems should be eliminated one by one, laying the road open for future users of data...

A lot of work ahead (2)

Experience has shown that the process of creating useful data sets takes time; it takes time to prepare the data, it takes time for modellers to use them, and it takes time to revise the data set in response to the feedback received.
All parties involved in evaluation activities must be aware of this nature of things.
We should build on the experiences of others, and this is a long, continuing process.

Ideas

- **Only half of Kincaid data has been distributed**
- **Prairie Grass is an obvious candidate for the Model Validation Kit**
- **Excel utilities could be added**
- **Utilities for the ASTM procedure could be enhanced.**
- **Establish collection of model evaluation results.**