THE MODEL EVALUATION TOOLS (MET): COMMUNITY TOOLS FOR FORECAST EVALUATION

Tressa L. Fowler

Barbara Brown, John Halley Gotway, Randy Bullock, Eric Gilleland, David Ahijevych, and Tara Jensen

June 2009

MET: A community tool

- The MET goal:

  To provide a set of forecast evaluation tools that is
  - "State-of-the-art"
  - Openly available
  - "Created" and enhanced by the community
    - Evaluation methods
    - Graphical methods

- *Community* includes diverse users
  - WRF model developers
  - Developmental Testbed Center (DTC)
  - University researchers
  - Operational centers

MET has nearly 500 registered users:
Roughly 50 / 50 %
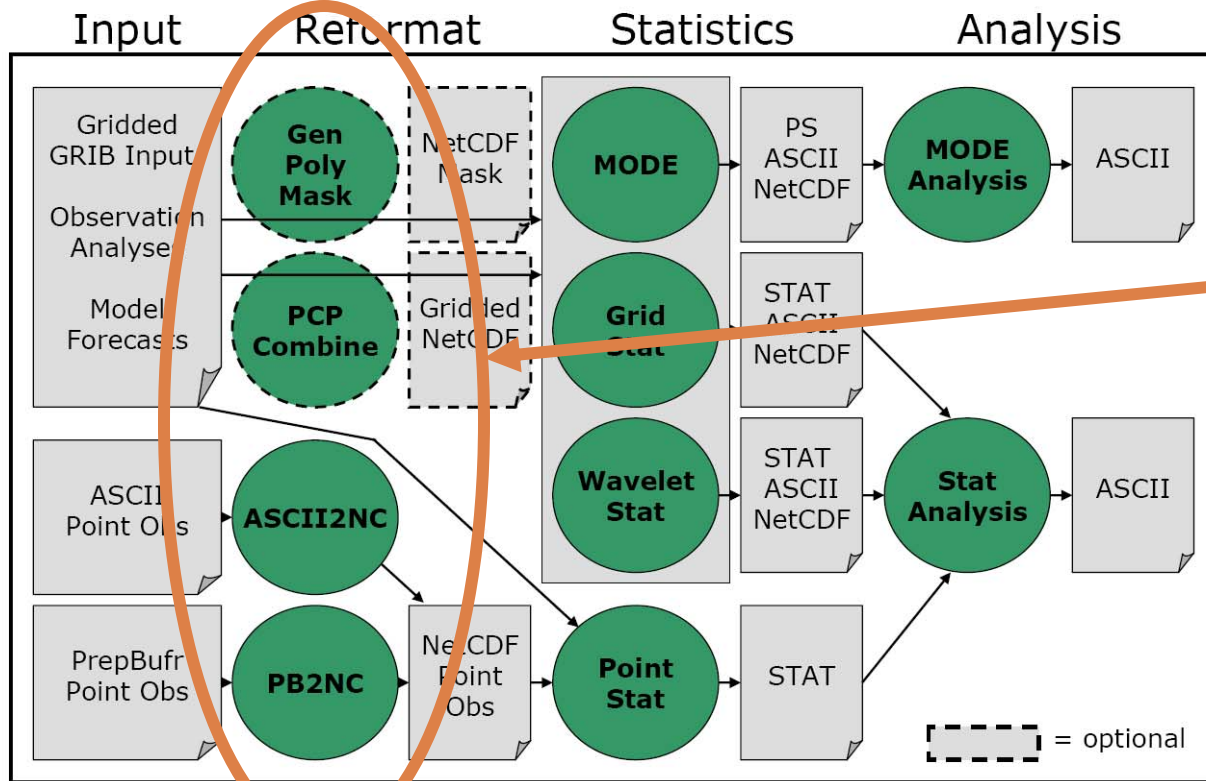University / Non-University

MET is…

- A modular set of forecast evaluation tools

- *Freely available*

- Highly *configurable*

- Fully documented

- Supported through the web and an e-mail help

MET is...

Reformatting tools:

Place data in the format(s) expected by the statistics tools

MET is...

MET v2.0 Flowchart
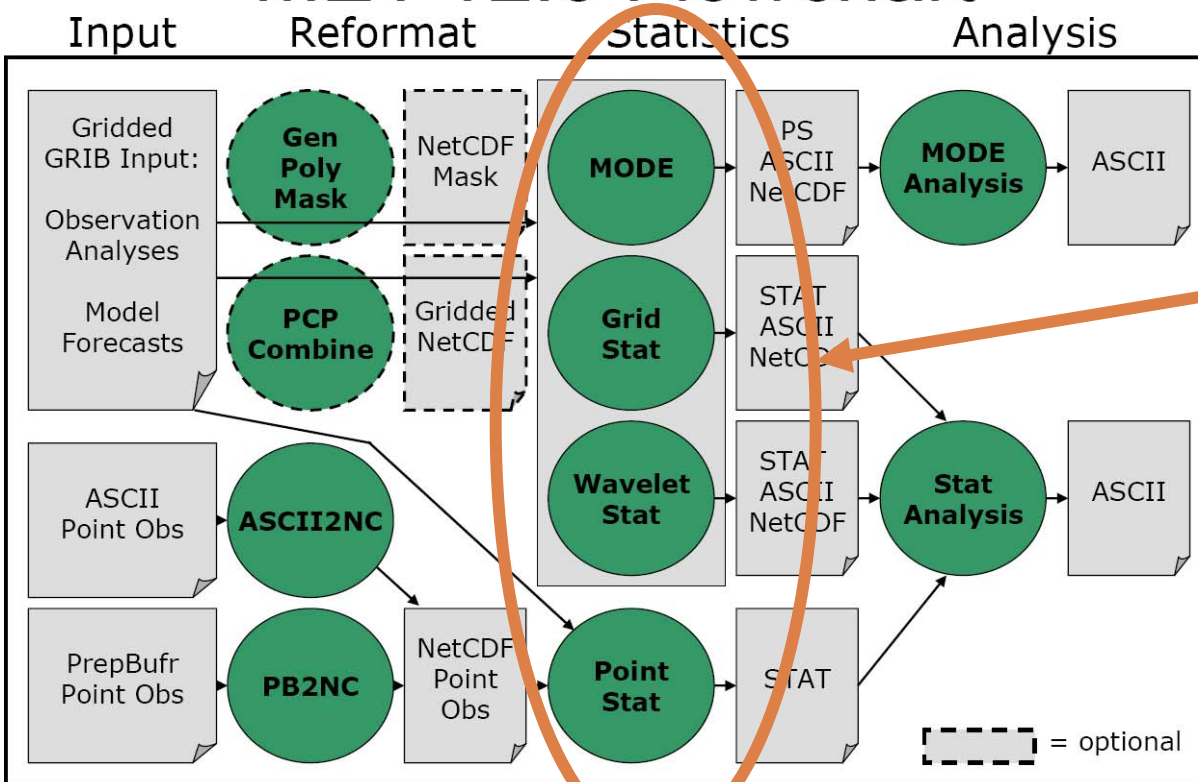
Statistics tools
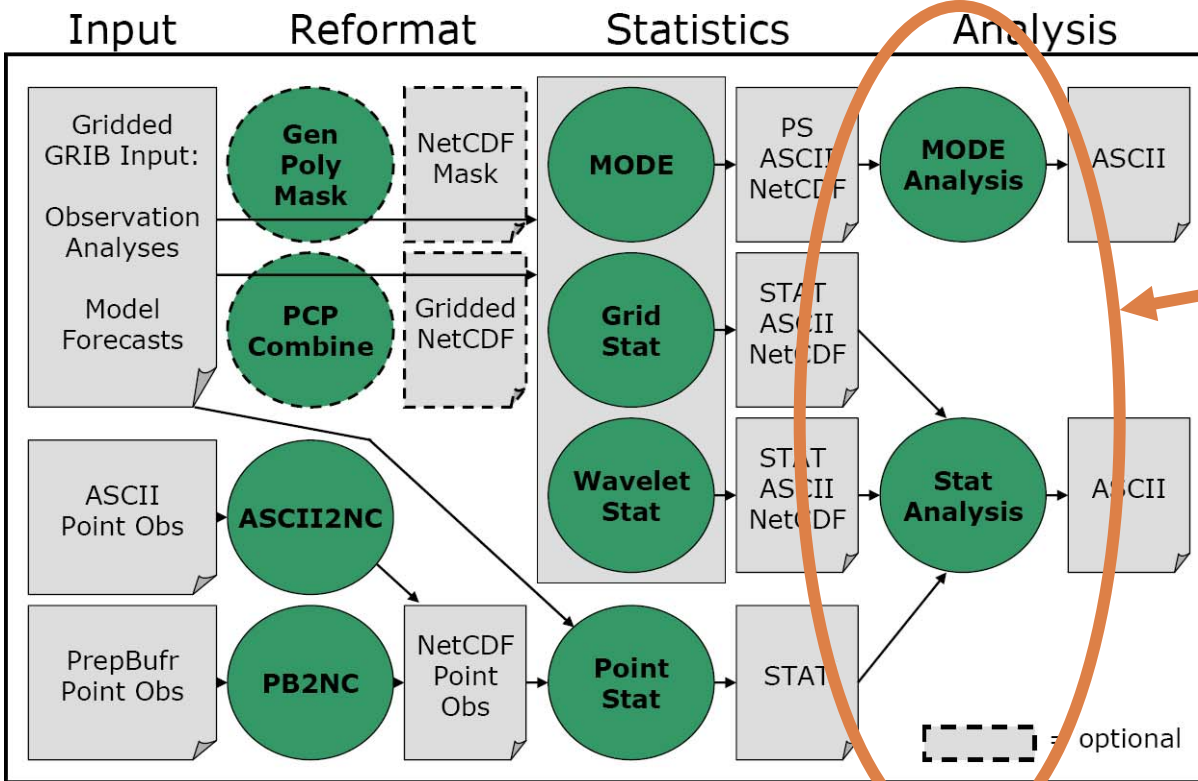
- Traditional methods
  - Gridded obs
  - Point obs
  - Confidence intervals

- Spatial methods
  - Object-based
  - Neighborhood
  - Wavelet (v2.0)

MET is…

MET v2.0 Flowchart

Analysis tools

□ Summarize statistics across cases

□ Stratify according to various criteria (e.g., lead time)

MET Statistics modules: Traditional verification measures

- Gridded and point verification
  - Multiple interpolation and matching options

- Statistics

  - **Continuous** - RMSE, BCRMSE, Bias, Correlation, etc.

  - **Categorical** - POD, FAR, CSI, GSS, Odds Ratio, etc.

  - **Probabilistic** - Brier Score, Reliability, ROC, etc. in v2.0

Matching approaches:

MET allows users to select the number of forecast grid points to match to a point observations and the statistic to use to summarize the forecasts.

MET Statistics modules: Confidence Intervals (CIs)

- MET provides two CI approaches
  - Normal
  - Bootstrap

- CIs are critical for appropriate and meaningful interpretation of verification results
  - Ex: *Regional comparisons*
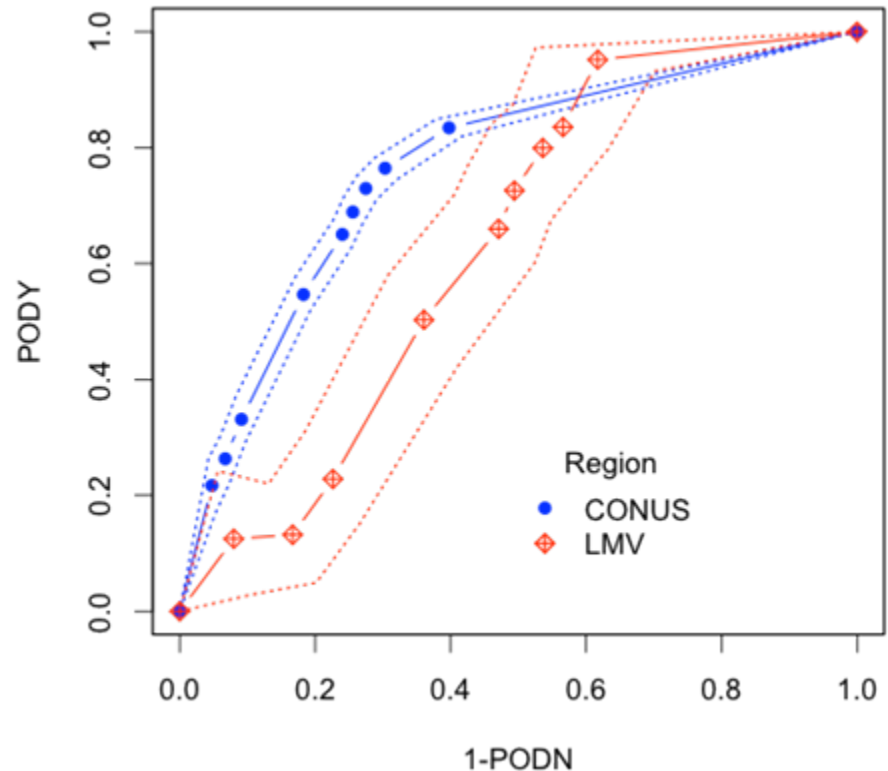
Verifying Probability Forecasts

- Probabilistic verification methods added to Grid-Stat, Point-Stat, and Stat-Analysis.
- Define Nx2 contingency table using:
  - Multiple forecast probability thresholds
  - One observation threshold
- Example:
  - Probability of precip [0.0, 0.25, 0.50, 0.75, 1.0]
  - Accumulated precip > 0.0

- Statistical Output:
  - Nx2 Table Counts
  - Joint/Conditional factorization table with calibration, refinement, likelihood, and base rate by threshold
  - Receiver Operating Characteristic (ROC) plot points by threshold
  - Reliability, resolution, uncertainty, area under ROC Curve, and Brier Score

Forecast	Observation		Total
	$o = 1$ (e.g., "Yes")	$o = 0$ (e.g., "No")	
p_1 = midpoint of (0 and threshold1)	n_{11}	n_{10}	$n_{1.} = n_{11} + n_{10}$
p_2 = midpoint of (threshold1 and threshold2)	n_{21}	n_{20}	$n_{2.} = n_{21} + n_{20}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_j = midpoint of (thresholdi and 1)	n_{j1}	n_{j0}	$n_{j} = n_{j1} + n_{j0}$
Total	$n_{.1} = \Sigma n_{i1}$	$n_{.0} = \Sigma n_{i0}$	$T = \Sigma n_{j}$

Simple ROC Plot Created Using MET Text Output

Receiver Operating Characteristic
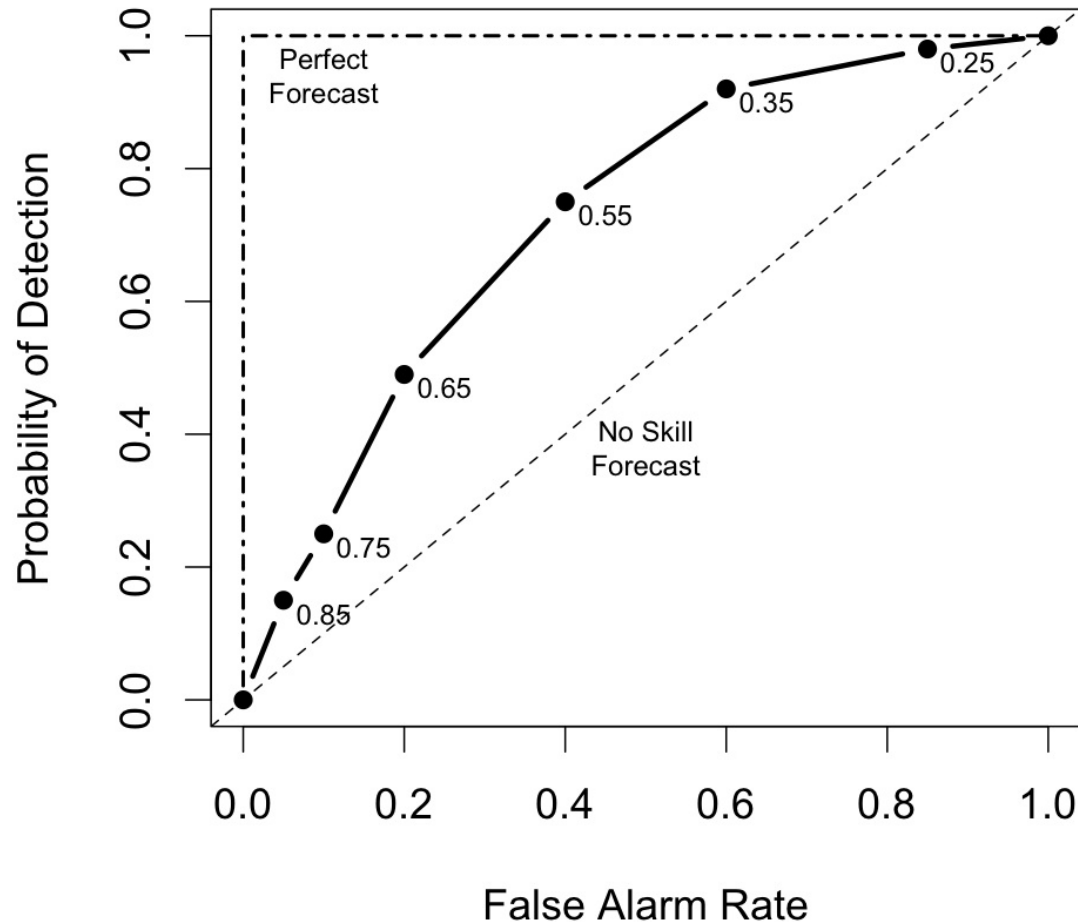
MET Statistics modules:
Spatial verification approaches

- Meaningful evaluations of spatially-coherent fields (e.g., precipitation)

  ### *Examples*

  - *What* is wrong with the forecast?

  - At what scales does the forecast perform well?

  - How does the forecast perform on attributes of interest to users?

- Methods included in MET

  - Object-based: Method for Object-based Diagnostic Evaluation (MODE)

  - Neighborhood; Example: Fractional Skill Score (FSS)

  - Scale-separation: Casati's Intensity-Scale measure (v2.0)

MODE Example

24-h precip forecast   Precip analysis
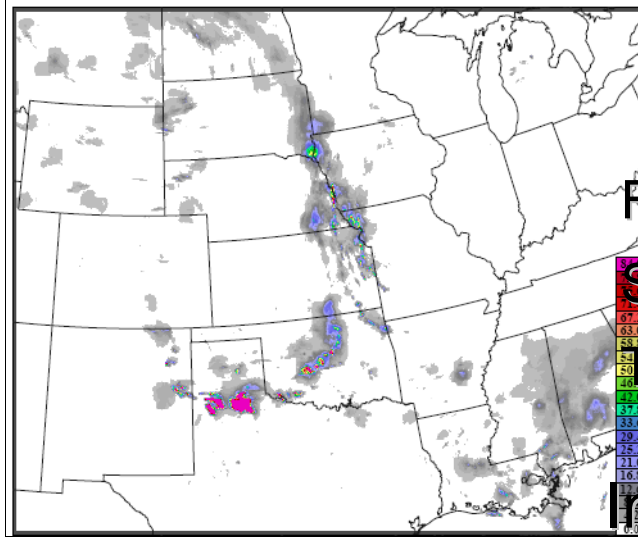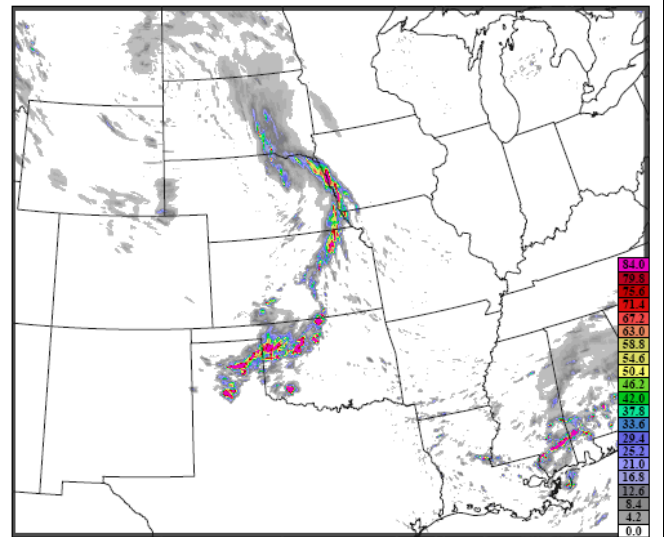
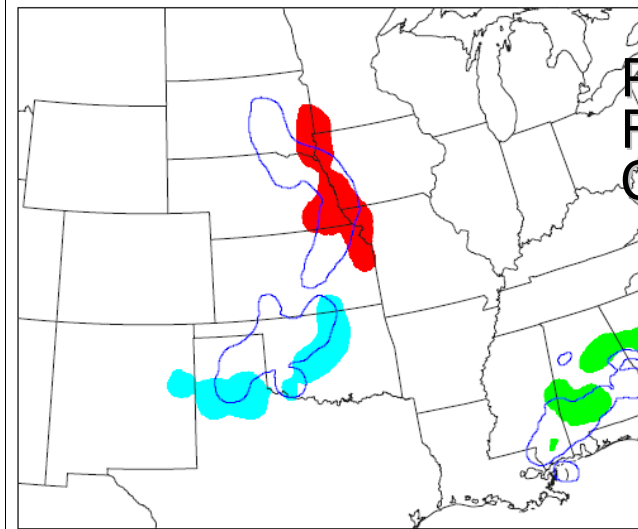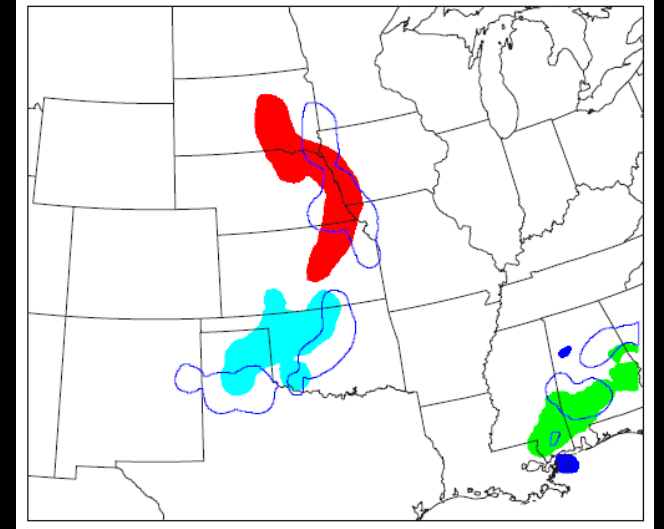MODE quantitative results indicate

Forecast is good

Slightly displaced

Too intense

In contrast:

POD = 0.40
FAR = 0.56
CSI = 0.27

Wavelet-Stat Tool

- Implements Intensity-Scale verification technique, Casati et al. (2004)
- Evaluate skill as a function of intensity and spatial scale of the error.
- Method:
  - Threshold raw forecast and observation to create binary images.
  - Decompose binary thresholded fields using wavelets (Haar as default).
  - For each scale, compute the Mean Squared Error (MSE) and Intensity Skill Score (ISS).
  - At what spatial scale is this forecast skillful?
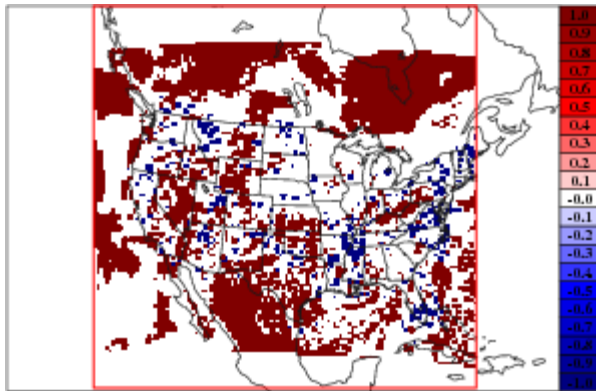
Difference (F-O) for precip > 0 mm
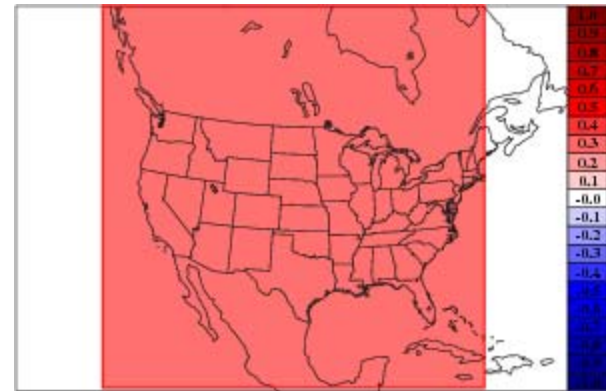
Wavelet decomposition difference

MET connections to the community
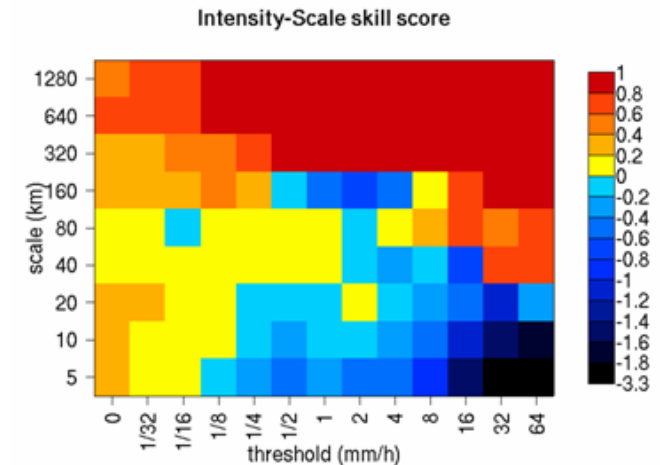
<u>Goals</u>:

Incorporate state-of-the-art methods contributed by the modeling, research, operational, and verification communities

Examples:
- Intensity-scale approach
- Neighborhood methods
- Graphical techniques

Outreach
- Town Hall meetings at AMS, NCAR
- Workshops (2007, 2008, 2009)
  - International verification experts + NWP experts + DTC staff
  - Guidance on methods and approaches to be included
- Spatial method intercomparison project (ICP)
- DTC Visitor Program
  - M. Baldwin: Verification testbed
  - B. Casati: Intensity-scale approach
- Demonstrations

Intensity-Scale skill score

Summary and plans

- MET is a community tool for forecast evaluation, which incorporates state-of-the-art methods
  - Modular architecture
  - Highly configurable
  - Extensive user support

- Plans and goals
  - Later versions
    - Ensemble forecasts, Cloud verification, Additional spatial methods, Wind methods
  - Database and display capabilities
  - Training
    - WRF tutorial (Feb. '09, July '09)
    - WRF Users' Workshop (June 2009)
  - Additional contributions from the community!
    - Tools
    - Graphics

For more information:

http://www.dtcenter.org/met/users/

MET Development Team

- Dave Ahijevych ⎫
- Tara Jensen ⎬ Scientists

- Barbara Brown ⎫
- Tressa Fowler ⎬ Statisticians/scientists
- Eric Gilleland ⎭

- Randy Bullock ⎫
- John Halley Gotway ⎬ Software engineers
- Steve Sullivan ⎭

For more information: http://www.dtcenter.org/met/users/