Model Evaluation Tools

# MET and Aerosols

**Subtitle:**
What would I do if it were me?

6th ICAP Workshop Oct 21-24, 2014 – Boulder, Colorado

Tara Jensen

MET Team: John Halley Gotway, Randy Bullock, Tressa Fowler, Tatiana Burek, Julie Prestopnik, Nancy Rehak, Kathryn Newman and Barb Brown

DTC Developmental Testbed Center

NCAR

# MET
Model Evaluation Tools

## Support for MET is provided by

AFWA, NOAA and NCAR

through the Developmental Testbed Center (DTC)

# Verification group vision

Provide verification expertise and software for NWP community to facilitate intelligent and defensible R2O decisions.

Reason to use MET: reproducible results across multiple institutions

MET
Model Evaluation Tools

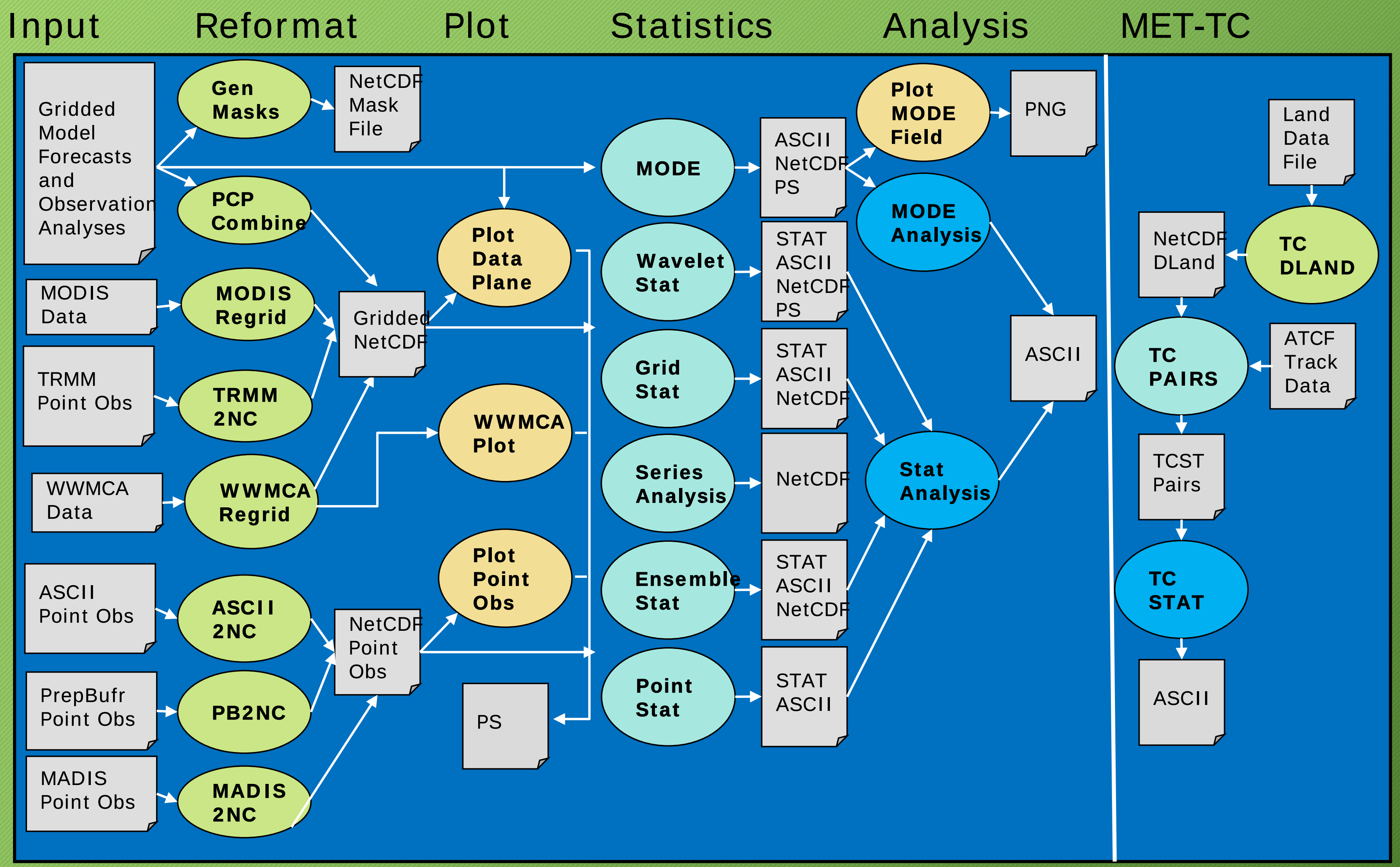
# MET Package

- MET is community code supported by DTC that is free to download (registration required)
  - Approximately 2500 registered users
  - 124 countries
  - Universities, Government, Private Companies, Non-Profits
- Download MET release and compile locally.
  - Register and download: **www.dtcenter.org/met/users**
- Language: Primarily in C++ with calls to some Fortran libraries
- Supported Platforms and Compilers:
  - Linux with GNU compilers
  - Linux with Portland Group (PGI) compilers
  - Linux with Intel compilers
- In-person tutorials given yearly – Next Tutorial – Feb 2-3, 2015 – after WRF Tutorial

# Things to keep in mind

- Things that MET does well:
  - Handles Grib1, Grib2, MET-Specific NetCDF, CF-compliant NetCDF, ascii
  - Calculates traditional and cutting edge statistics for gridded datasets against either point data or gridded data
  - Reproduces and extends
    - Basic NCEP FVS verification package
    - NHC Verification Package
- Things that MET is still working on:
  - Handling HDF files (support needed)
  - Handling 2 “point” fields
  - Providing pre-compiled binaries
  - Verification against aircraft data

# MET v5.0 Overview



| Data | MET Tool |
| --- | --- |
| Gridded Forecasts<br>Gridded Observations<br><br>***(Grib1 / Grid2 / NetCDF with grid information included and next release GSI diagnostic file)*** | Grid Stat (traditional or neighborhood)<br>Ensemble Stat<br>Wavelet Stat<br>MODE<br>Series Analysis<br>Stat Analysis |
| Gridded Forecasts<br>Point Observations<br><br>***(Ascii in 11 column format; PrepBufr)*** | Point Stat<br>Ensemble Stat<br>Stat Analysis |
| Point Forecasts<br>Point Observations<br><br>***(ATCF format)*** | TC Pairs<br>TC Stat |

**MET has finally broken the requirement for gridded data paradigm!!**

# Categorical and Continuous Statistics

| Continuous | Categorical / Multi-Categorical |
|---|---|
| Forecast Mean<br>Forecast Standard Deviation<br>Observation Mean<br>Observation Standard Deviation<br>Pearson Correlation Coefficient (aka Correlation)<br>Spearman's Rank Correlation<br>Kendall's Tau statistic<br>Number of ranks used in Kendall's tau<br>Number of tied forecasts in Kendall's tau<br>Number of tied observations in Kendall's tau<br>Mean error<br>Standard Deviation of error<br>$10^{th}$, $25^{th}$, $50^{th}$, $75^{th}$, $90^{th}$ Percentile of Error<br>Inner Quartile Range<br>Multiplicative Bias (aka Bias)<br>Mean Absolute Error<br>Mean Square Error<br>Bias-corrected Mean Square Error<br>Root Mean Square Error<br>Mean Absolute Deviation<br><br>24 Statistics | Total number of matched pairs<br>Contingency Table Counts<br>Forecast rate<br>Hit rate<br>Observation rate<br>Base rate<br>Forecast mean<br>Accuracy<br>Frequency Bias<br>Probability of Detection - Yes<br>Probability of Detection - No<br>Probability of False Detection (aka False Alarm Rate)<br>False Alarm Ratio<br>Critical Success Score (aka Threat Score)<br>Gilbert Skill Score (aka Equitable Threat Score)<br>Bias-Adjusted Gilbert Skill Score<br>Odds Ratio<br>Log-Odds Ratio<br>Odds-Ratio Skill Score<br>Hanssen-Kuipers Discriminant<br>Heidke Skill Score<br>Extreme Dependency Score<br>Symmetric Extreme Dependency Score<br>Extreme Dependency Index<br>Symmetric Extreme Dependency Index<br><br>25 Statistics |

# Neighborhood and Ensemble/Probability Statistics

RUN:

GRID-STAT

RUN:

WAVELET-STAT

| Neighborhood | Ensemble/Probability |
|---|---|
| Neighborhood Contingency Table Statistics<br>(see previous slide)<br><br>Fractions Brier Score<br>Fraction Skill Score<br>Asymptotic Fractions Skill Score<br>Uniform Fractions Skill Score<br>Forecast Event Frequency<br>Observed Event Frequency | Ensemble Mean and Std Dev fields<br>Ensemble Mean + 1 Std Dev fields<br>Ensemble Min and Max fields<br>Ensemble Range field<br>Ensemble Valid Data Count field<br>Ensemble Relative Frequency (probability)<br><br>Ranked Histograms (if Obs Field Provided)<br>PIT Historgram<br>Ensemble Spread-Skill (if Obs Field Provided) |
| **Wavelet** | Neighborhood Contingency Table Statistics<br>(see previous slide)<br>Brier Score<br>Reliability<br>Resolution<br>Uncertainty<br>Area Under ROC<br>Calibration<br>Refinement<br>Likelihood<br>Base Rate<br>Probailiby Integral Transform (PIT)<br>Reliability points<br>ROC Curve Points |
| Mean squared error for each scale<br>Intensity skill score<br>Forecast Energy Squared<br>Observed Energy Squared<br>Base Rate (not scale dependent)<br>Frequency Bias | |

RUN:

ENSEMBLE STAT

RUN:

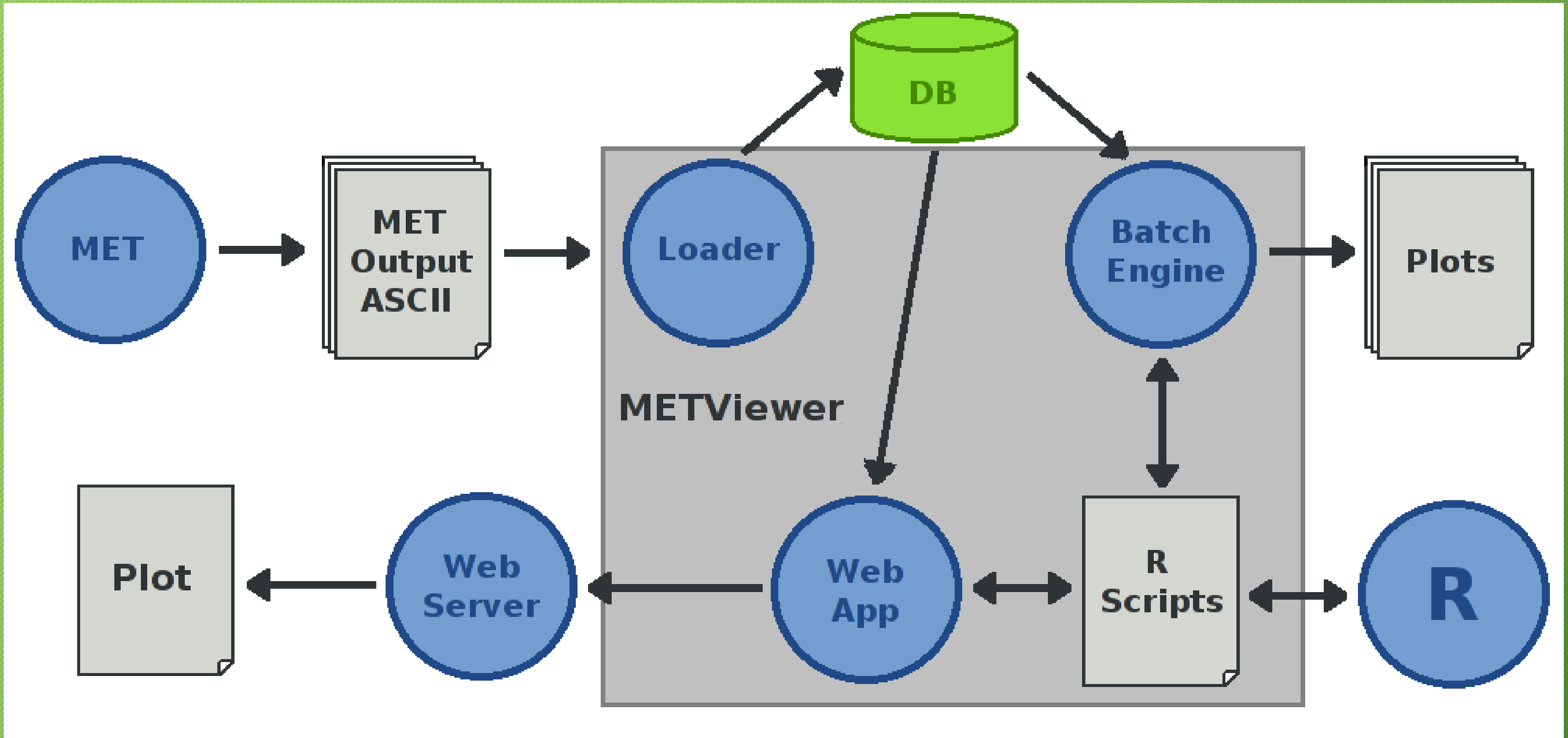
POINT-STAT OR GRID-STAT

# Analysis and Synthesis of Statistics

- Stat-Analysis
- MODE Analysis
- Series Analysis
- MET-TC TC-Stat
- METViewer Database and Display package

*(currently available through DTC – will work with people to upload some of their data so you can figure out if it's a useful tool you want to adopt)*

# METViewer Design



# METViewer 1.0 Release

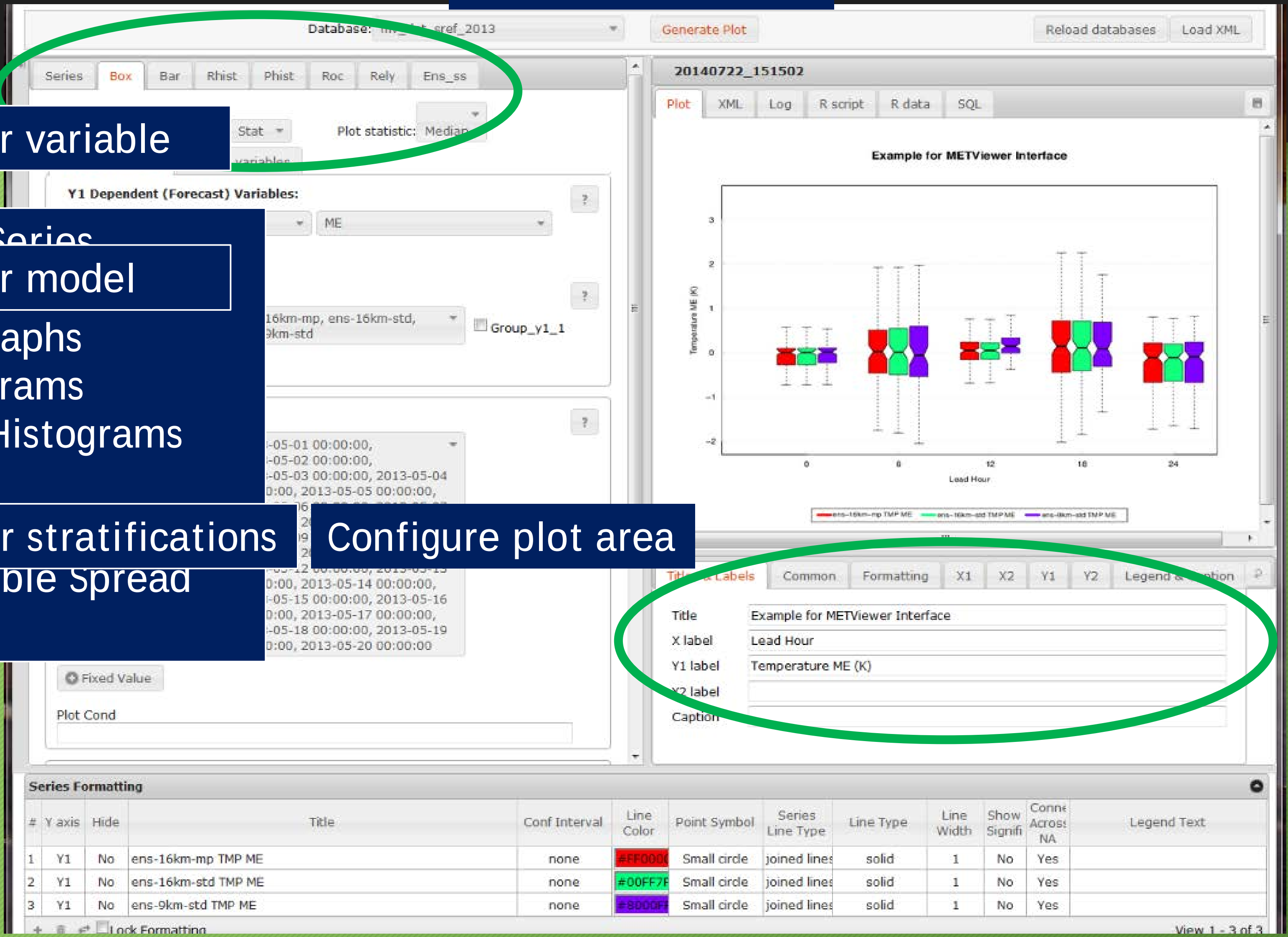
***User interface redesigned***
***Functionality the same***
***Regression Testing Implemented***

**Reasons to use METViewer**

- Easy way to analyze data
- Allows you to aggregate either by finding the mean or median of errors – or - using partial sums (for continuous variables) and contingency tables (for categorical data)
- Pairwise differencing
- Confidence intervals
- Plot formatting on the fly
- Batch engine to produce multitudes of plots
- We now have regression testing for METViewer

# METViewer 1.0

Many plot options




Pick your variable

Time Series

Pick your model

Bar Graphs
Histograms
Rank Histograms
ROC

Pick your stratifications

Configure plot area

Ensemble Spread Skill

# ICAP Examples

ME


Think both inside and outside the box

YOU

Slide Courtesy of Angela Benedetti

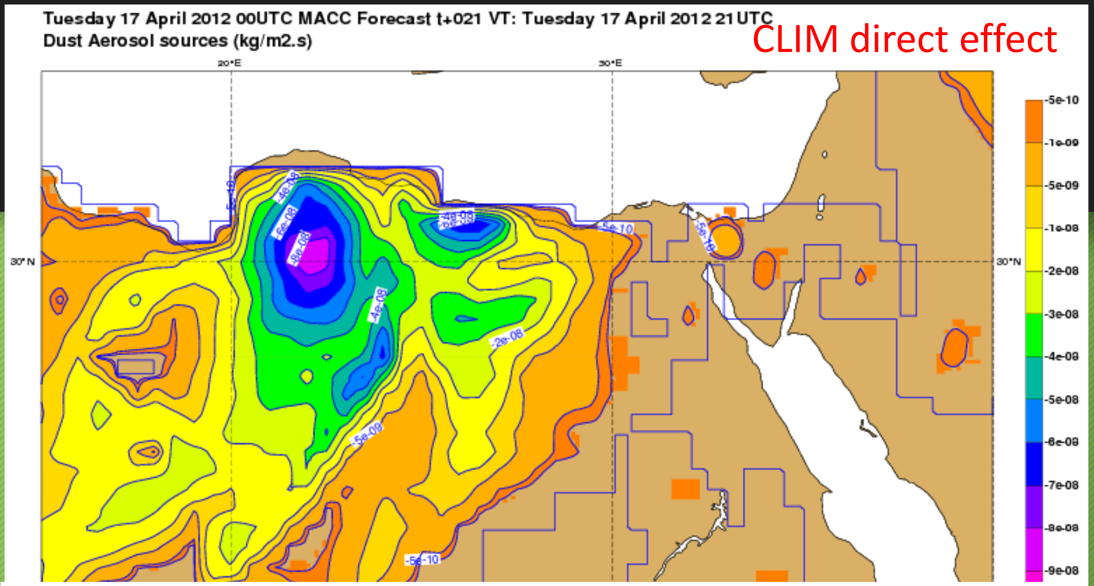
# AEROSOL RADIATIVE IMPACT ON NWP: sources comparisons

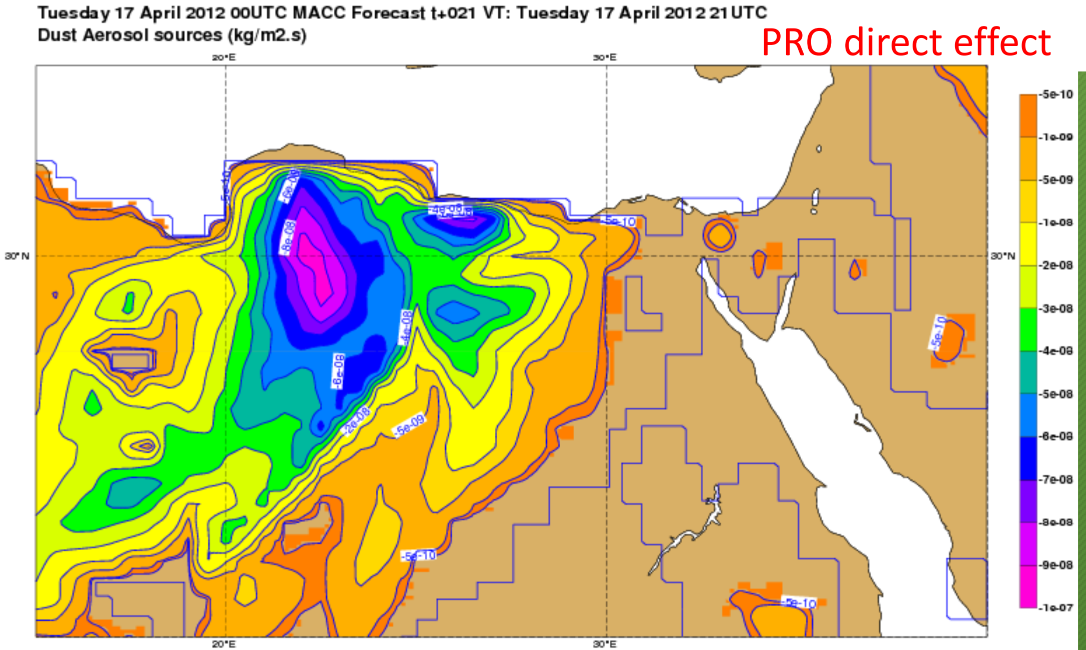
Aerosol (ie dust) sources are much larger with prognostic aerosol direct effect

I would use:

Grid-Stat to calculate categorical statistics from contingency statistics

MODE to better understand the "features"



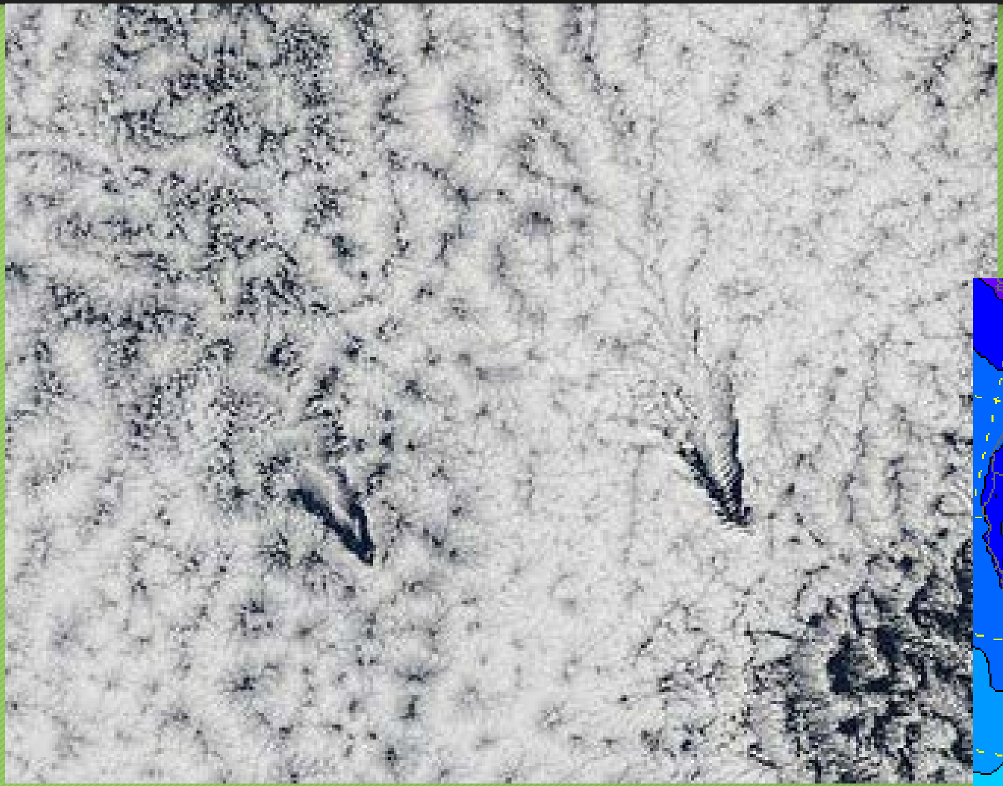

Samuel Remy

# Comparing objects can tell you things about your forecast like . . .

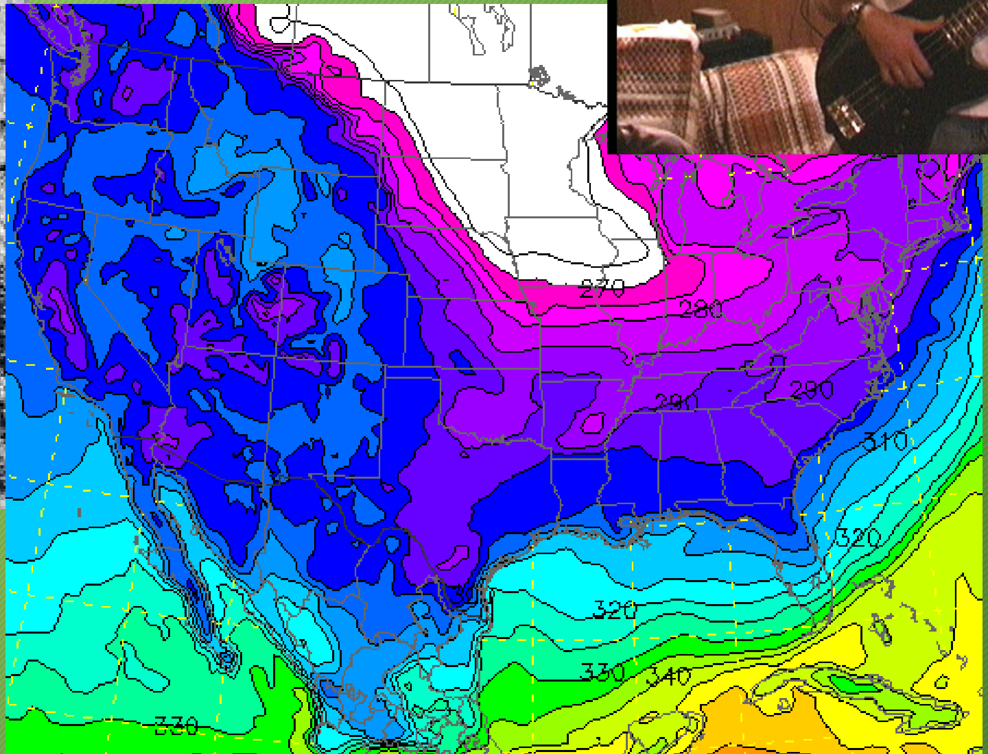
| This: | Instead of this: |
|---|---|
| 30% Too Big<br>(area ratio=1.3) | POD = 0.35 |
| Shifted west 1 km<br>(centroid distance = 1km) | FAR = 0.7235 |
| Rotated 15°<br>(angle diff = 15%) | CSI = 0.1587 |
| Peak AOD 50% too low<br>(diff in 90$^{th}$ percentile of intensities = 0.5) | |

IMAGES OF THE DAY

# Verifying with objects doesn't always make sense . . .



NASA/MODIS - **May 9, 2011 - Ship-wave-shape wave clouds induced by Juan Fernandez Islands, off Chile**




Slide Courtesy of Angela Benedetti

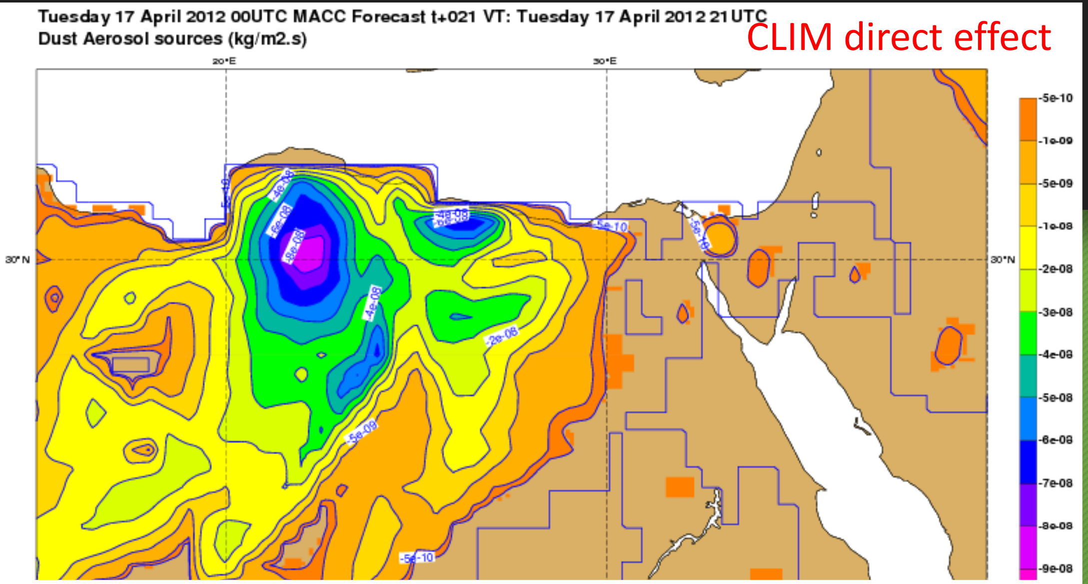
# AEROSOL RADIATIVE IMPACT ON NWP: sources comparisons

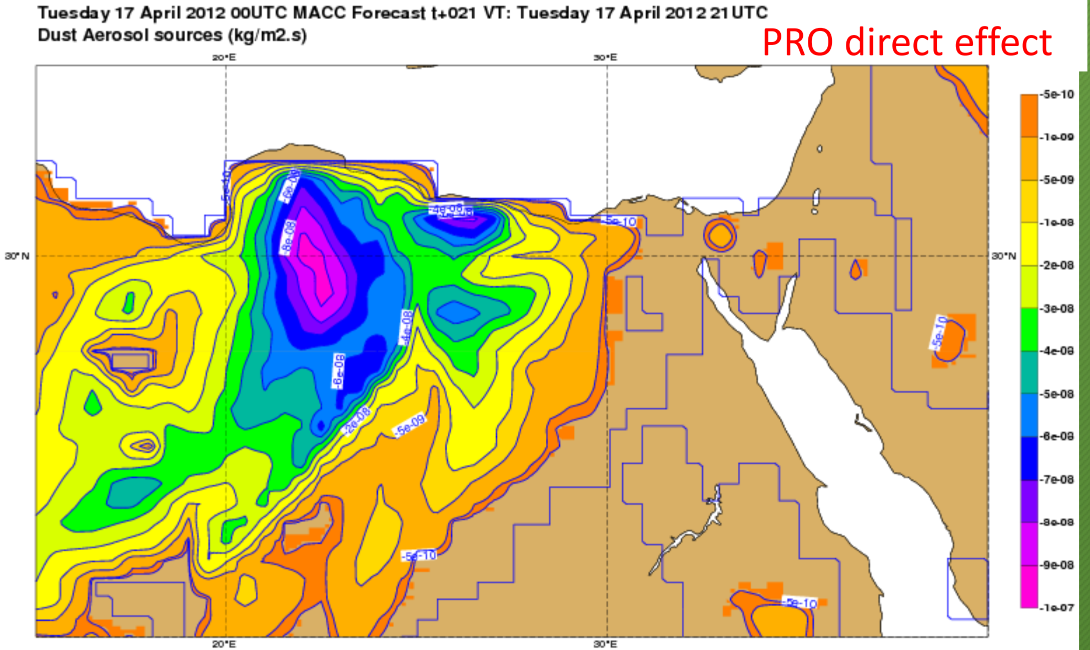
Aerosol (ie dust) sources are much larger with prognostic aerosol direct effect

I would use:

Grid-Stat to calculate categorical statistics from contingency statistics

MODE to better understand the "features"



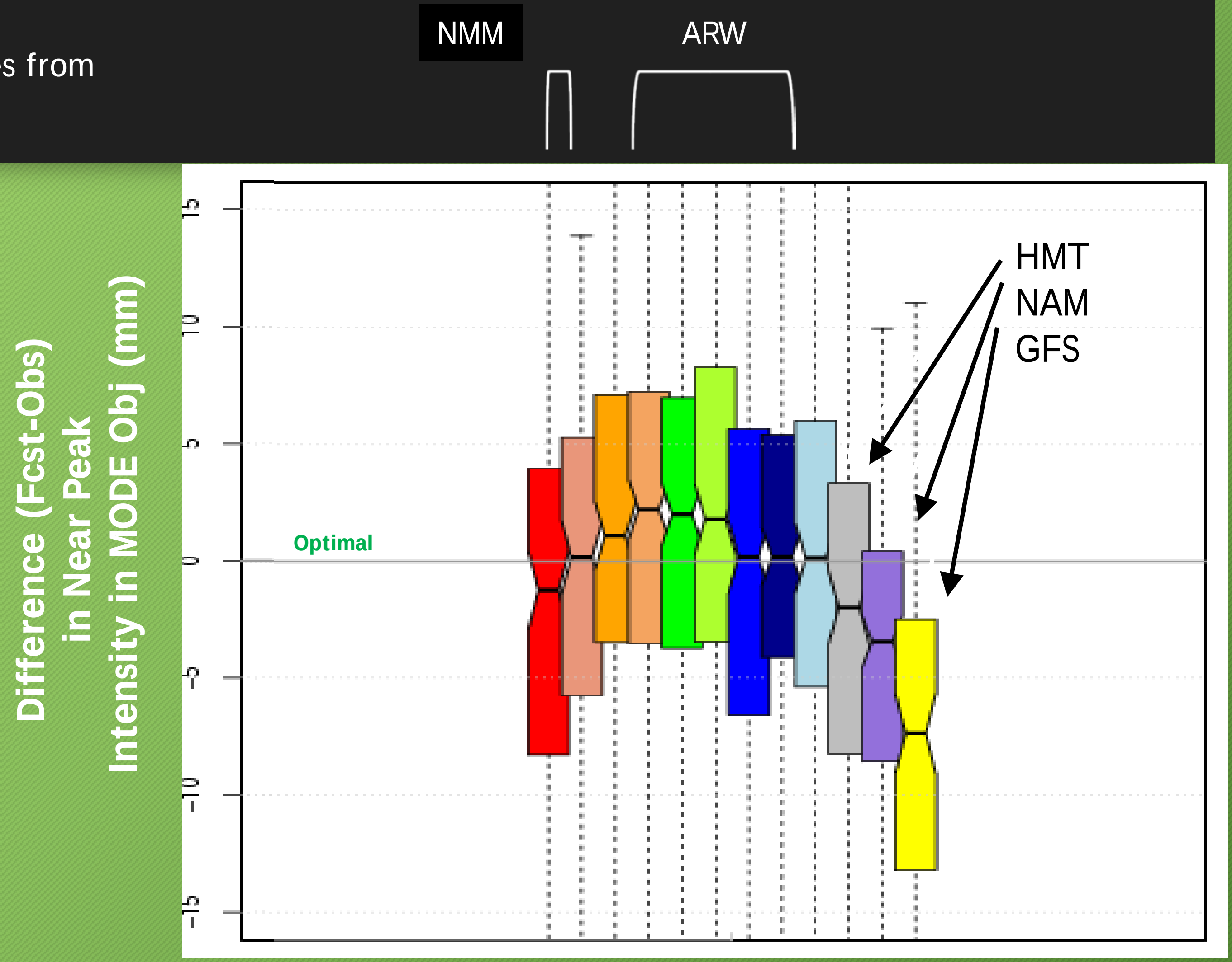




Samuel Remy

# Impact of Core & Microphysics

Using Attributes from
MODE Objects

## Intense Precipitation Cores

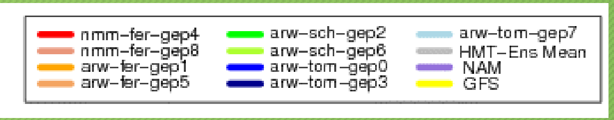



> 25.4mm
Object Threshold (mm)

**90% Intensity shows** over-forecast of precipitation for ARW-Fer and ARW-MY members especially at higher thresholds – which means **when it rains it pours in these members**

Color groups
Different
Microphysics

Ferrier
Milbrant-Yau
Thompson



# MODE has been used to evaluate

- Precipitation
  - Standard Accumulation Intervals
  - Probability Fields
- Reflectivity
  - Composite
  - Radar Echo Top
- Precipitable Water / IWV
- Cloud free areas
- A-Train 2-D vertical curtain of satellite fields
- World-Wide Merged Cloud Analysis (WWMCA)
- Mid-latitude Jets
- Confluence and Difluence derived from Wind Fields
- High and Low pressures

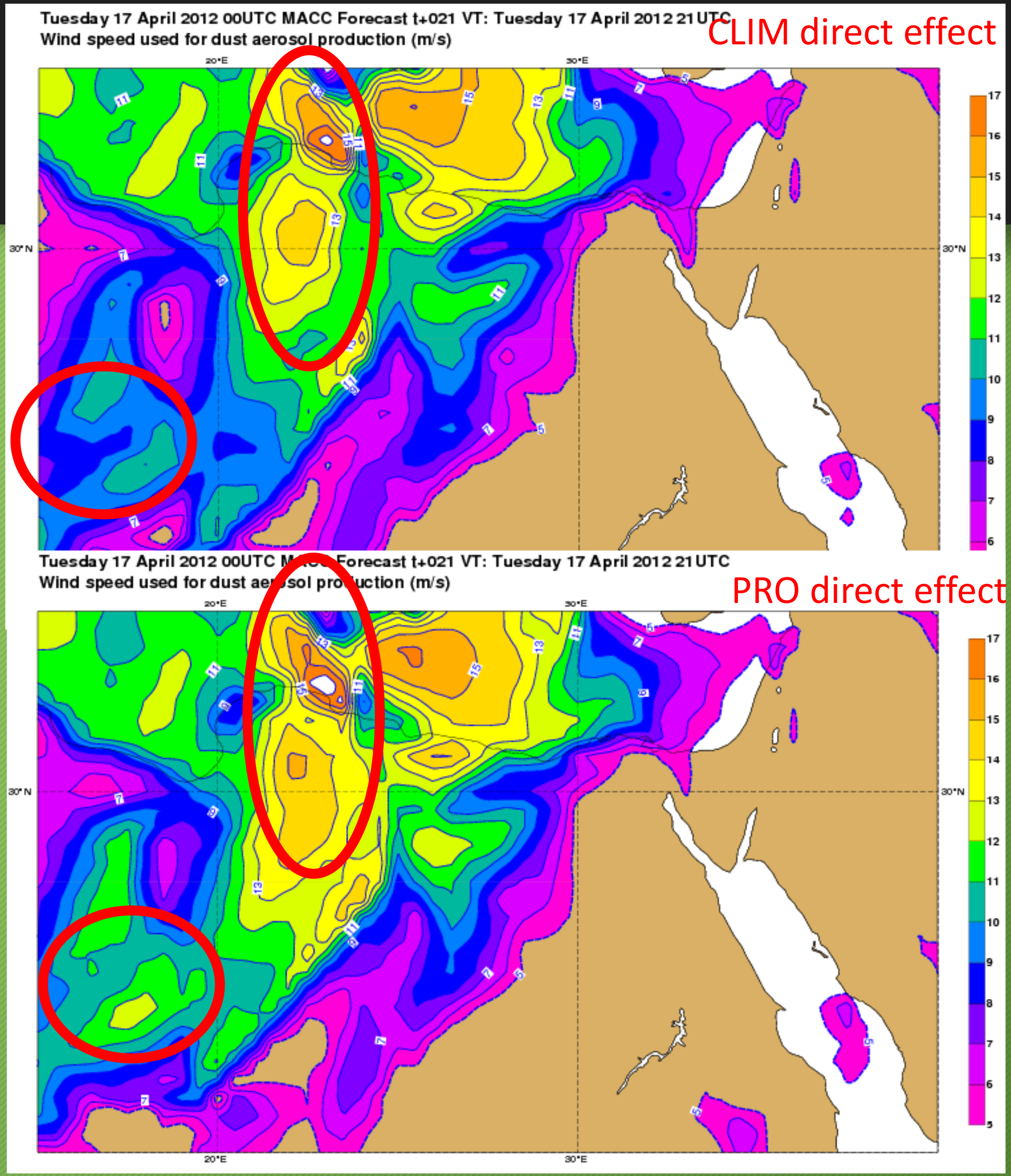
Slide Courtesy of Angela Benedetti

# AEROSOL RADIATIVE IMPACT ON NWP: 10m wind comparisons

I would use:

Grid-Stat to calculate continuous statistics and categorical statistics from contingency statistics

MODE to better understand the “features”

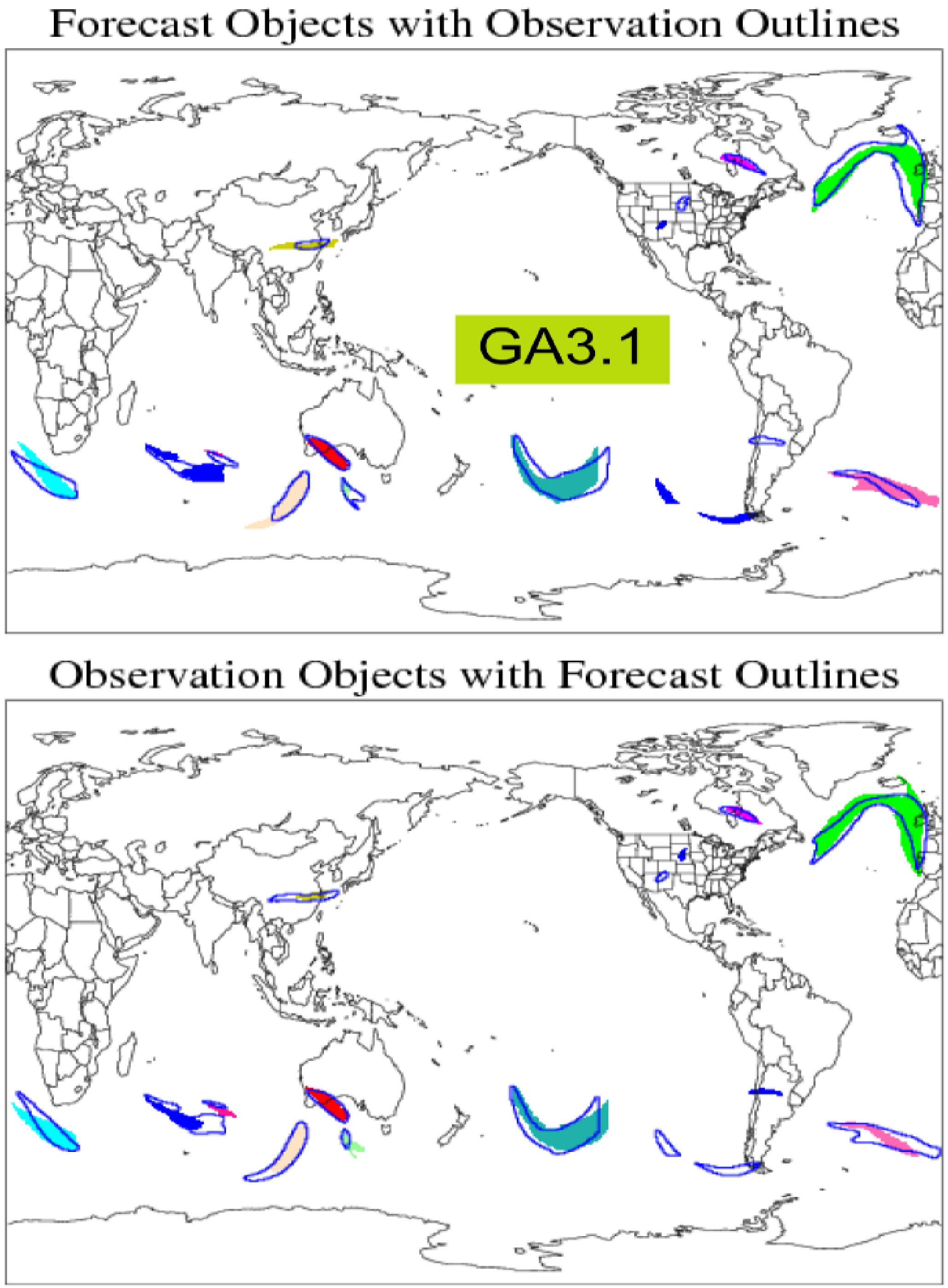



Samuel Remy

# Temporal evolution

- Older N320 trial
  250 hPa winds > 60 m/s
  at forecast lead time of t+96h
  from the 12Z initialisation
  compared to EC analyses

- Differences in the size of forecast and analysed objects is not overshadowed by growth of synoptic forecast error.




Example of MODE

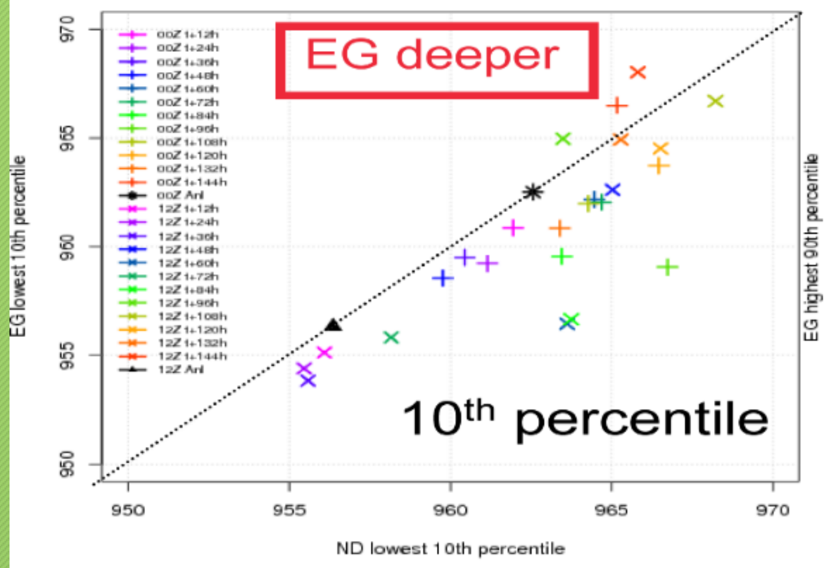
# Example of MODE Attributes

## Object intensities

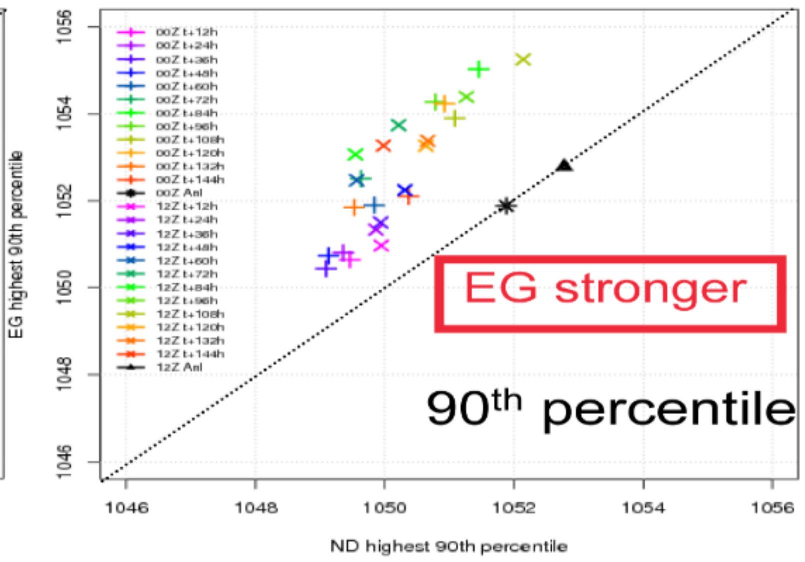
Met Office

EC analyses
N768 EG v N512 ND

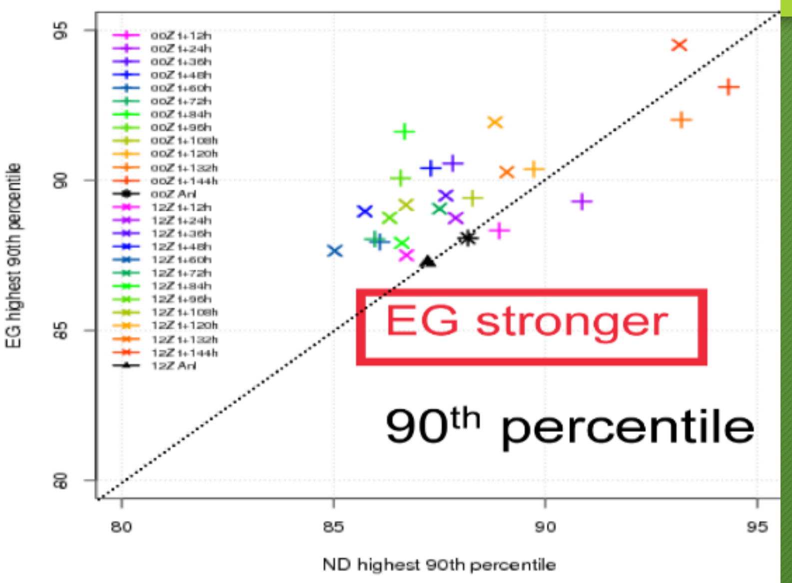
**Lows** NHW <=990mb




**Highs** NHW >=1036mb




**Jets** NHW >=60ms−1




- Do not look at absolute min/max values in objects. Use the 10th or 90th percentile as a more reliable estimate of how the intensity distribution has shifted/changed.
- Lows are deeper, highs and jets are stronger → sharper gradients and a more active energetic model.
- Differences in the 00Z and 12Z analyses.



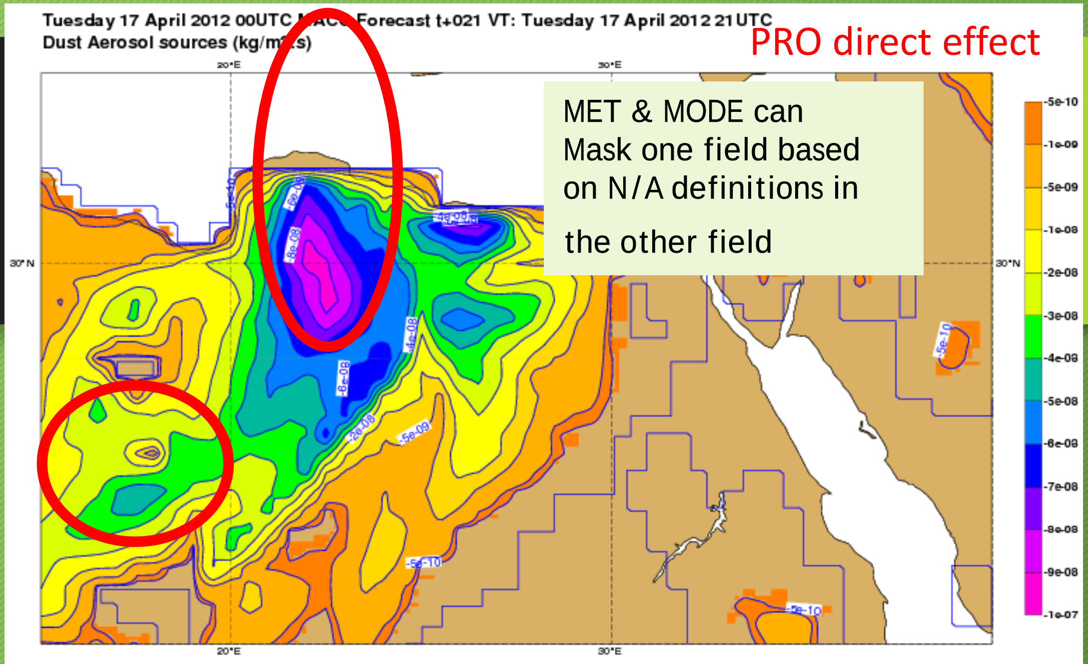
*Slide courtesy of Marion Mittermaier*
*The Met Office – August 2014*

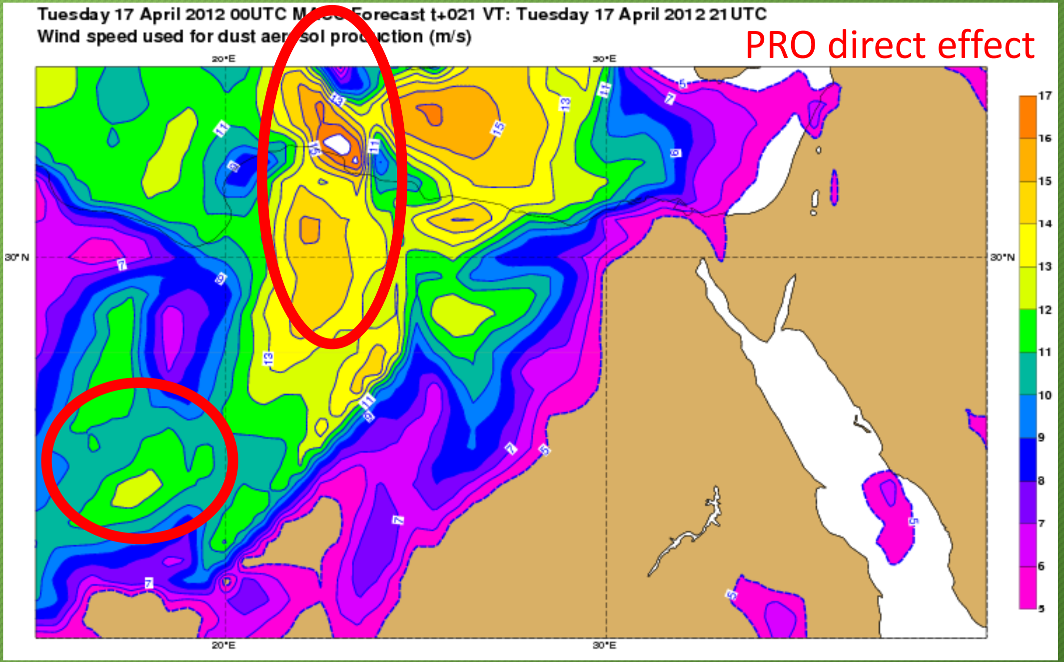
Slide Courtesy of Angela Benedetti

# AEROSOL RADIATIVE IMPACT ON NWP:
# Source and 10m wind comparisons

You could also use MODE to compare the two fields by thresholding

1) Source field with whatever is meaningful in that field and
2) Different winds thresholds to identify what it takes to forcing dust production in a give region

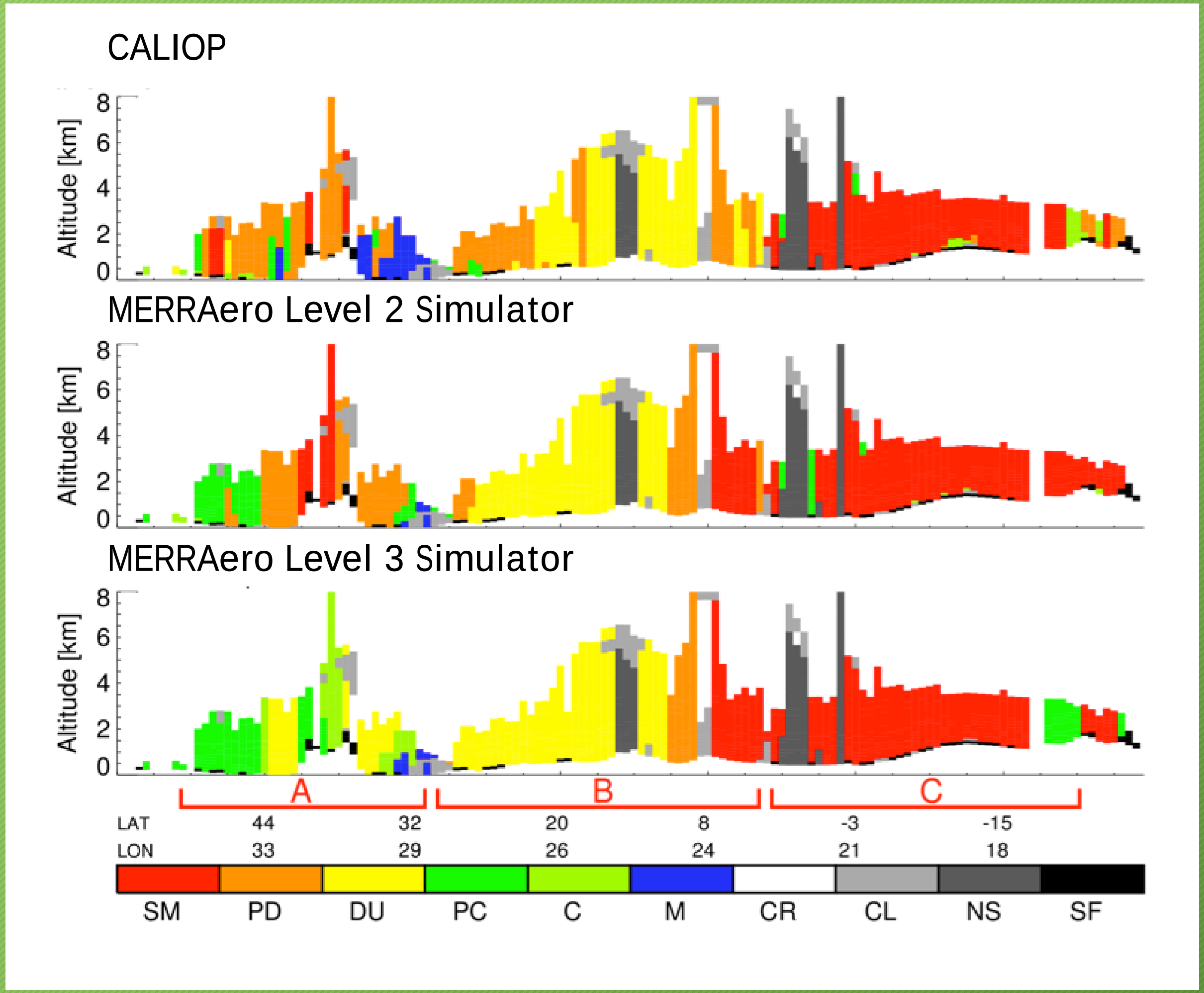



Samuel Remy

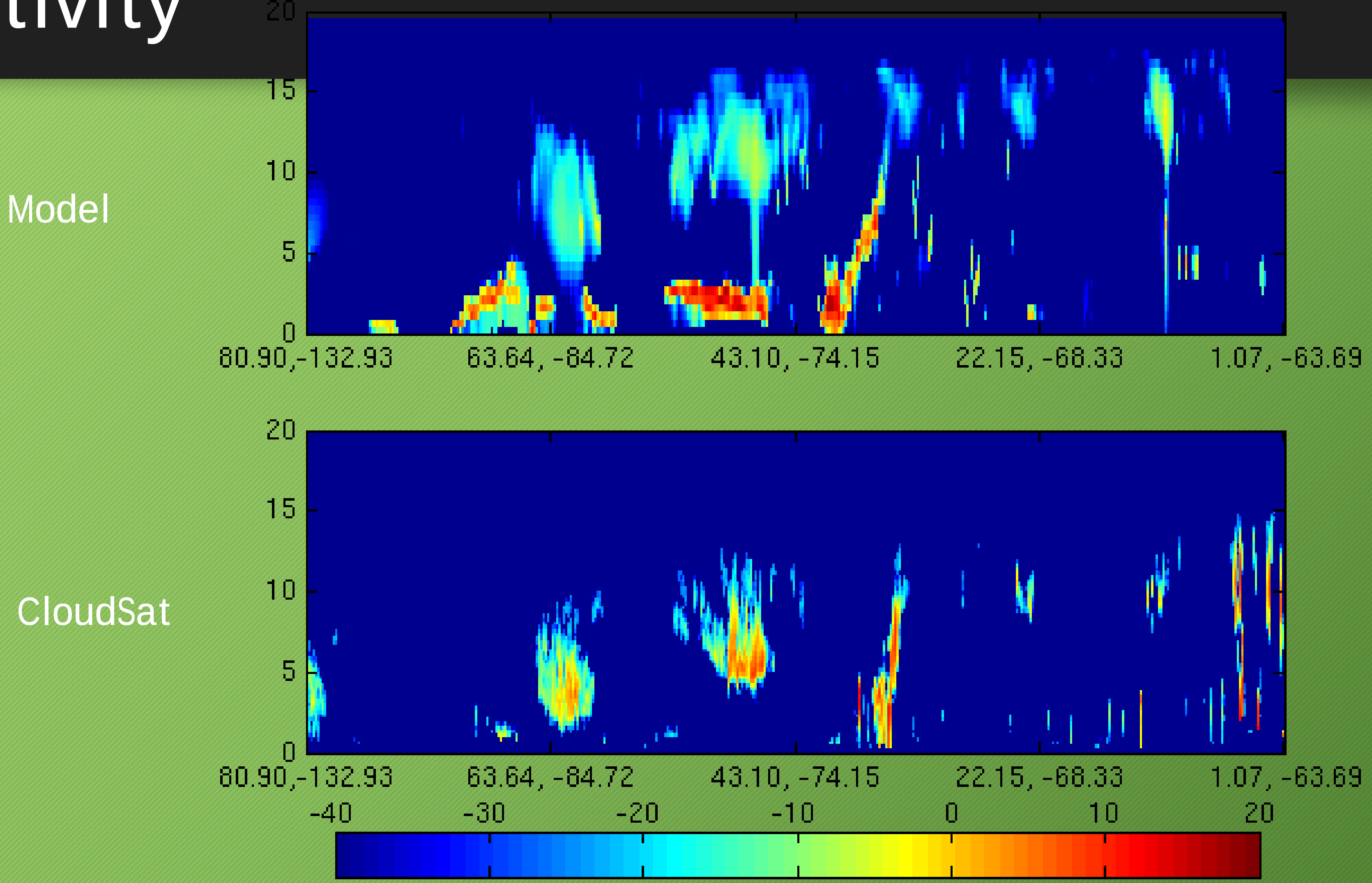
Slide Courtesy of
Peter Colarco

# MERRAero: Lidar Simulation




Evaluation of MERRAero aerosol typing with the CALIOP Vertical Feature Mask

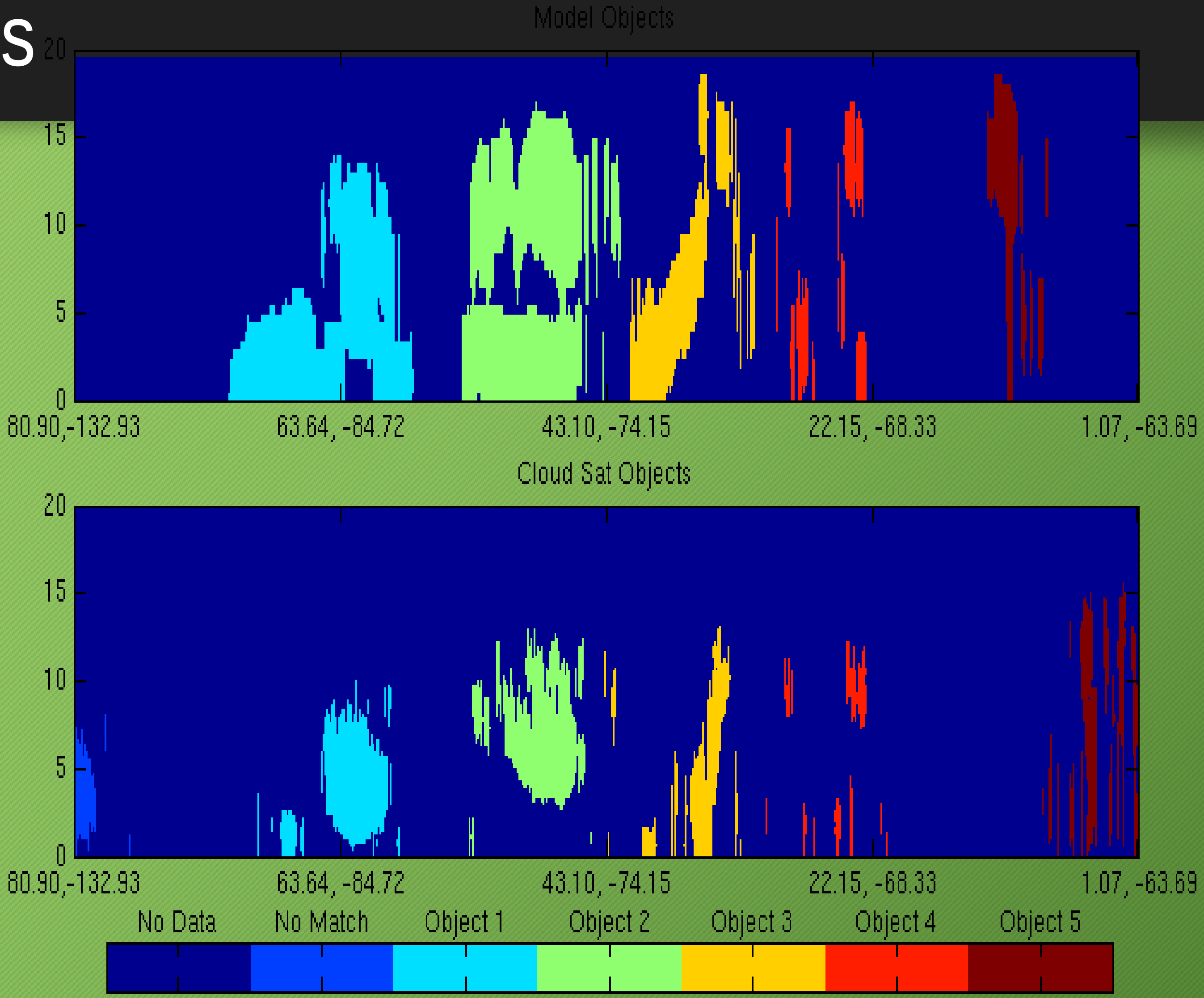
# CloudSat/NWP Comparison: Object Based: Reflectivity



# CloudSat/NWP Comparison: MODE Objects

Model

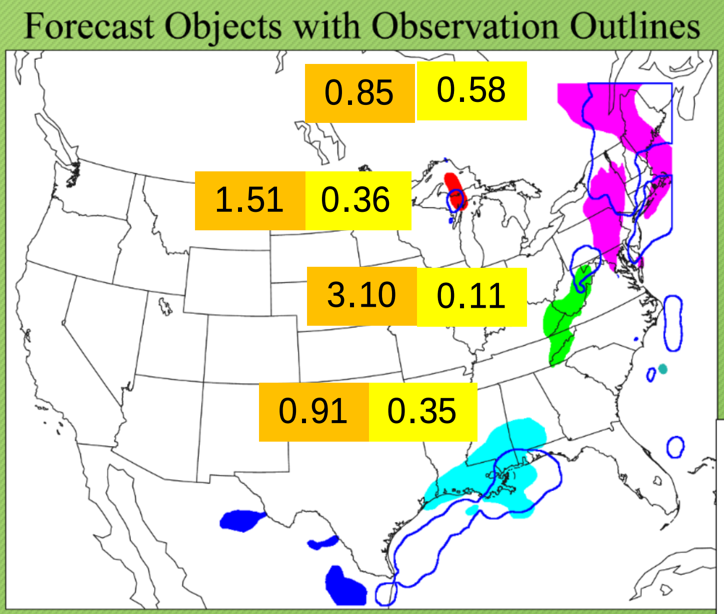
CloudSat




You could use MODE in vertical to calculate MODE attributes In the verticl plane

# MODE for Different Probabilities – May 11, 2013 (DTC SREF tests)

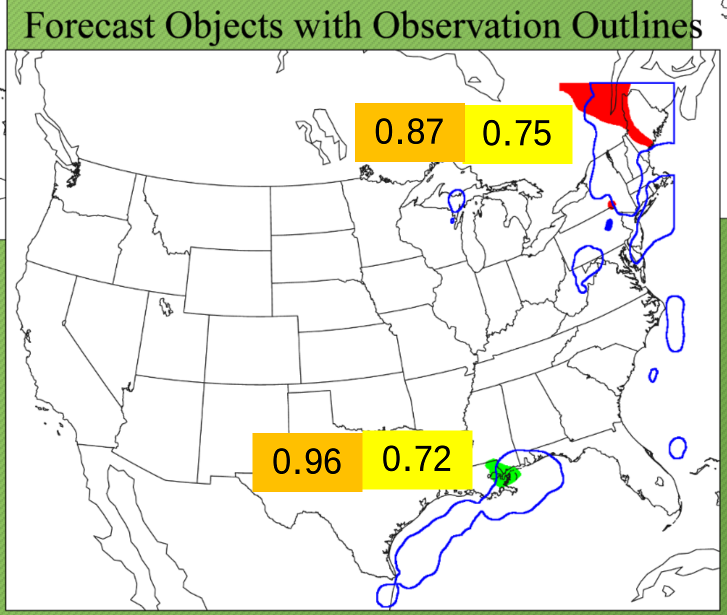
NWS PoP - Percent chance that rain will occur at any given point in the area.



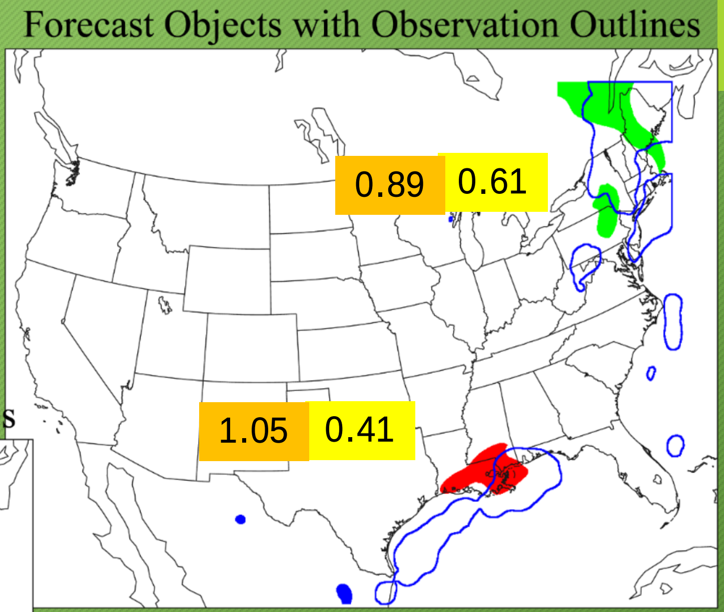

**Prob >25%** of 2.54mm in 3hr

A-Percent Area: **Intersection Area / Forecast Area**

C-Confidence: **Symmetric Difference** *(non-intersecting area)* **/ Observed Area**




**Prob >75%** of 2.54mm in 3hr



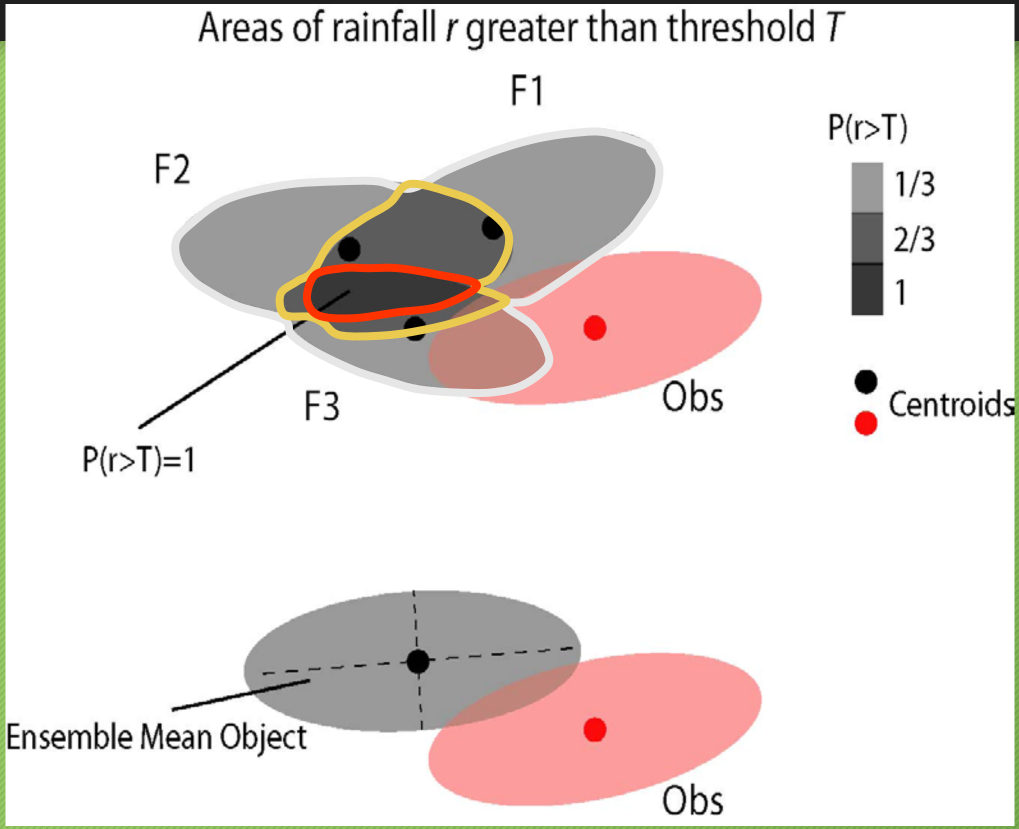

**Prob >50%** of 2.54mm in 3hr

**Observation**

**Forecast**

***NWS PoP*** *= C x A where*
*"C" = the confidence that precipitation will occur* ***somewhere*** *in the forecast area*
*"A" = the percent of the area that will receive measureable precipitation.*

# Applying spatial methods to ensembles




**As probabilities:** Areas do not have "shape" of precipitation areas; may "spread" the area

**As mean:**

Area is not equivalent to any of the underlying ensemble members

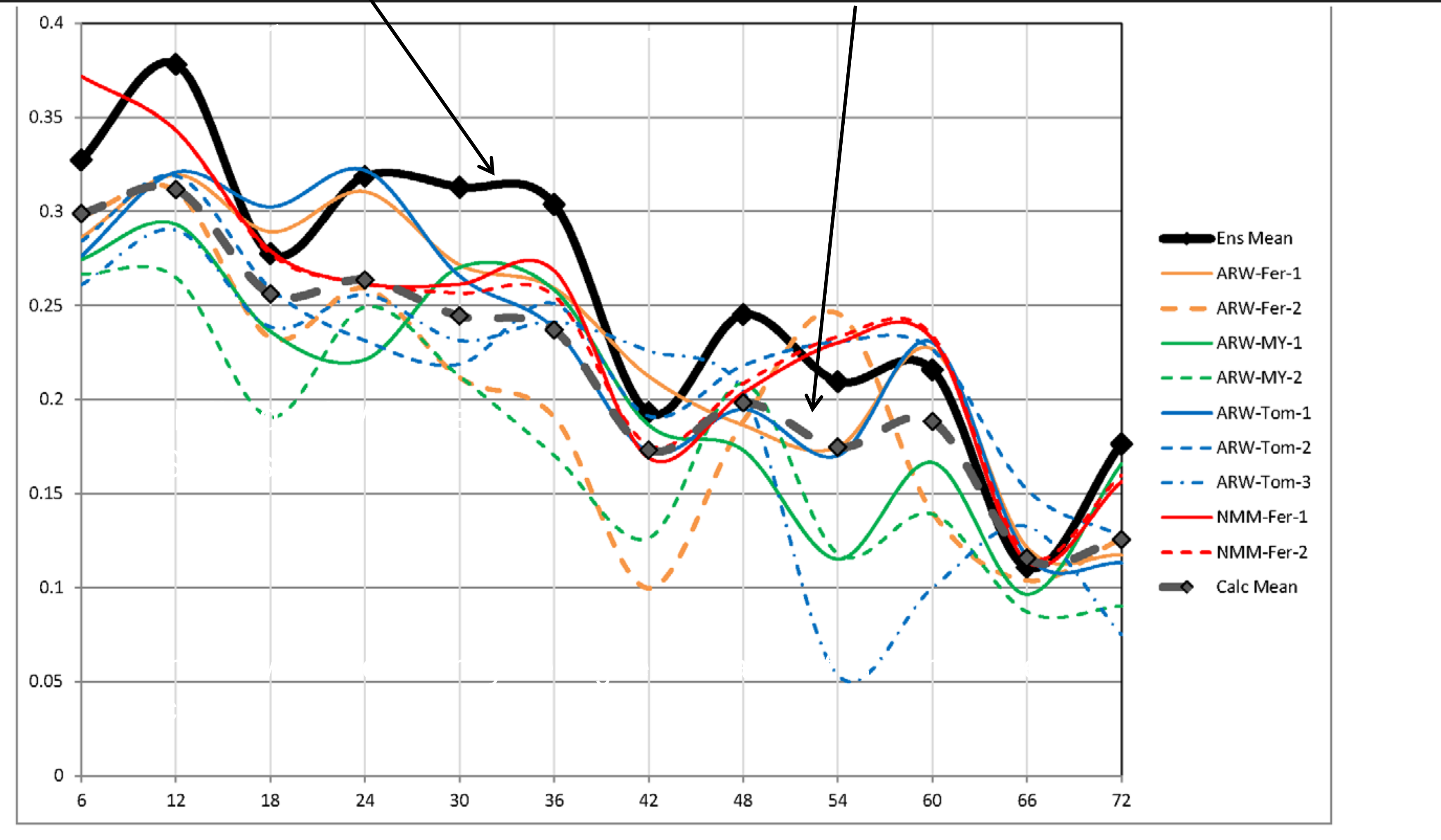
**As an ensemble of attributes:**

May have many interesting features

# Score of Ensemble Mean is not the Mean of the Ensemble Member Scores

**Ensemble Mean Score**

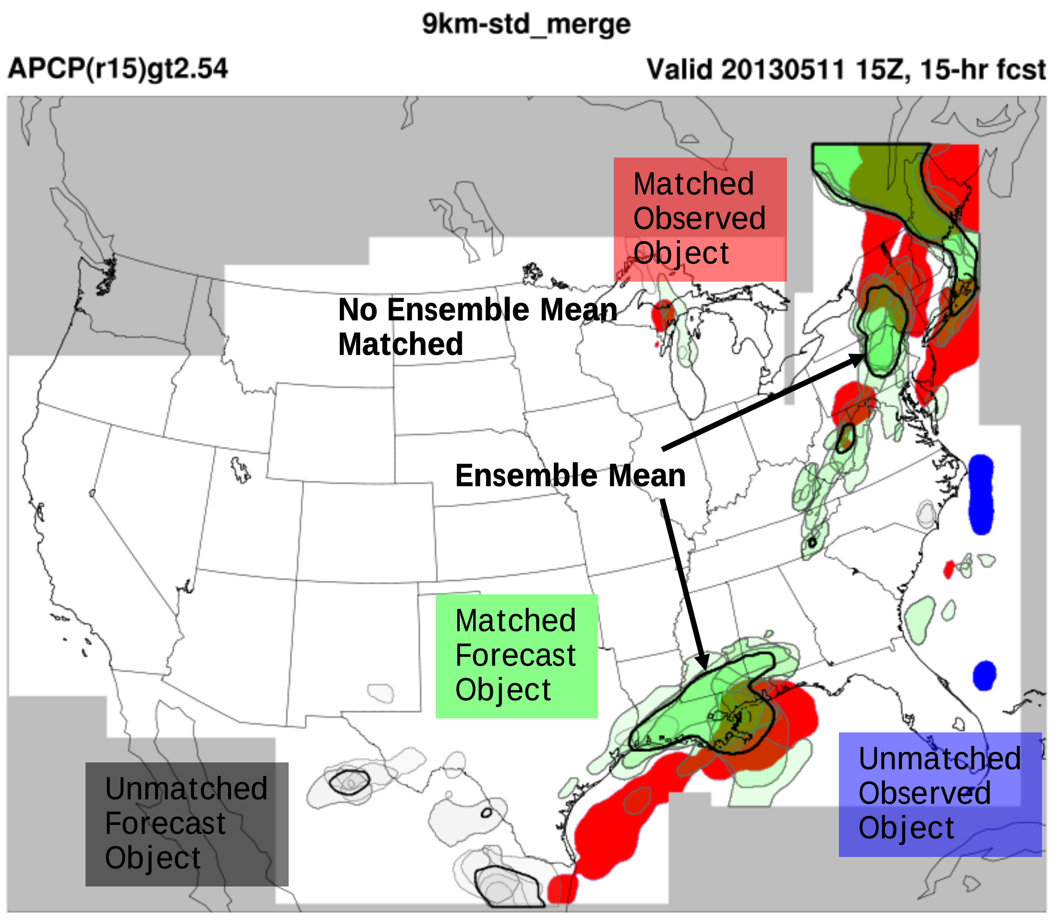
**Mean of Ensemble Member Scores**




DTC Developmental Testbed Center

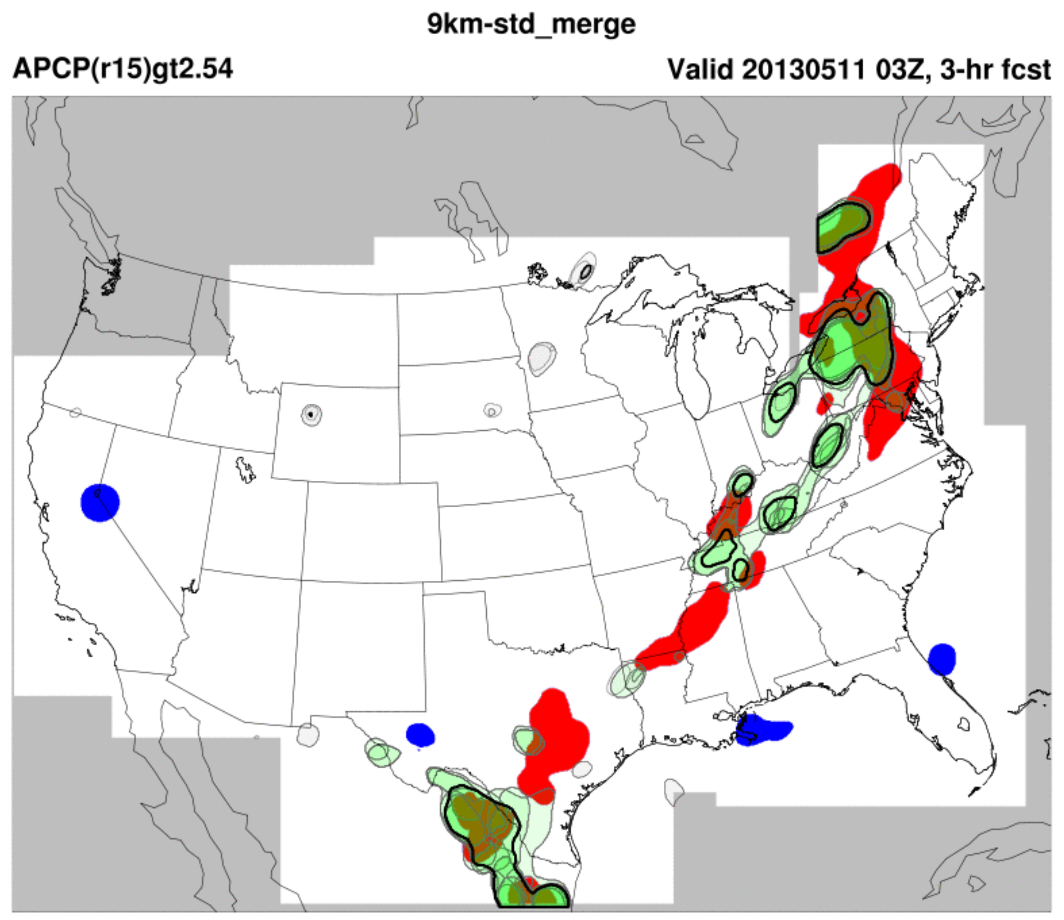
# Example May 11, 2013

# DTC SREF Tests – ARW Members




DTC Developmental Testbed Center

# Spread increases With Time




DTC Developmental Testbed Center

Slide Courtesy of Angela Benedetti

# PM VERIFICATION

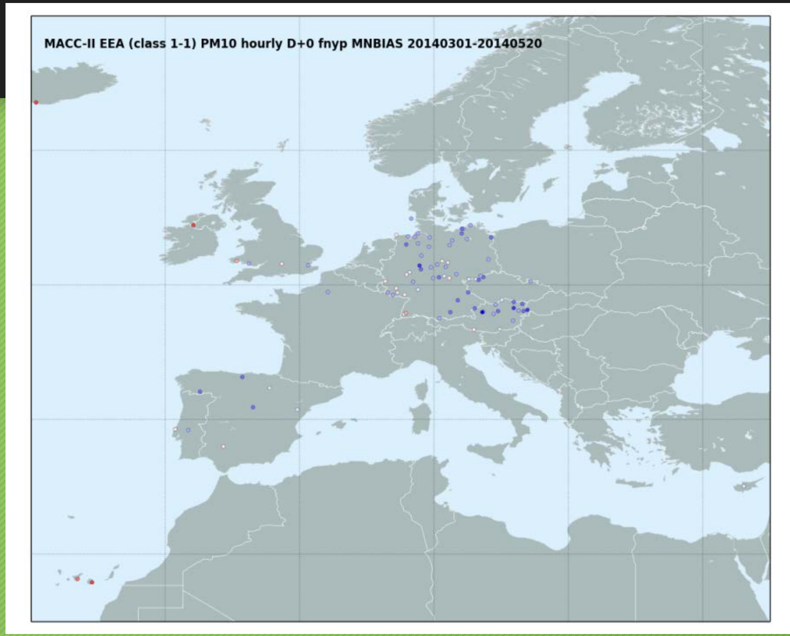
Miha Razinger

I would use:
Point-Stat with probably bilinear mean interpolation to match forecast with obs

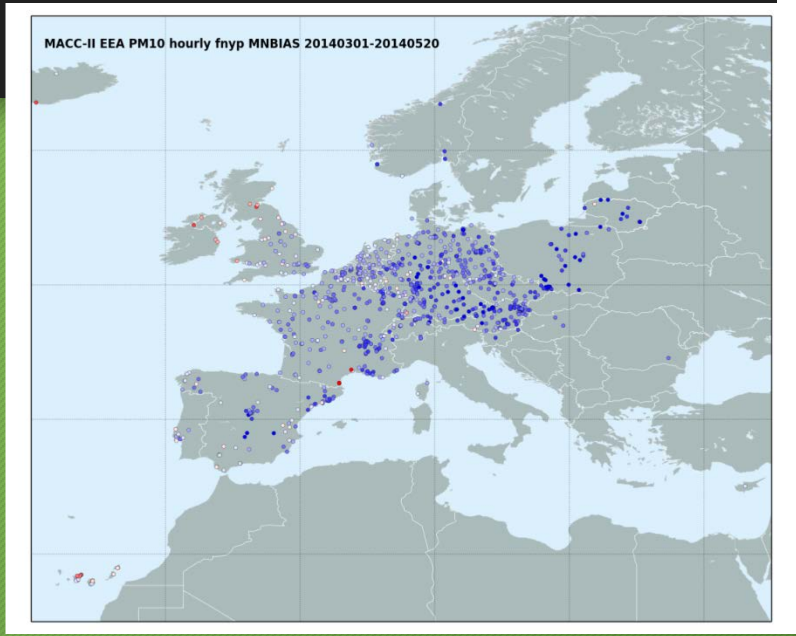
Stat-analysis to aggregate statistics over either:
- Sites
- Entire set of stations
- Regional masking regions

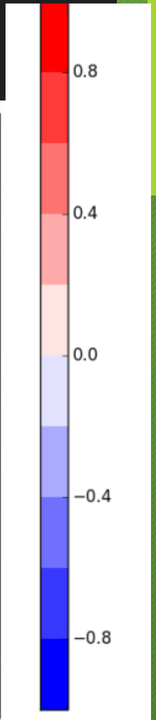
$$\text{Normalized Mean Bias} = \frac{2}{N}\sum_{i=1}^{N}\left(\frac{P_i - O_i}{P_i + O_i}\right)$$

**Class 1 Joly-Peuch classification= background stations**




**All reporting stations**



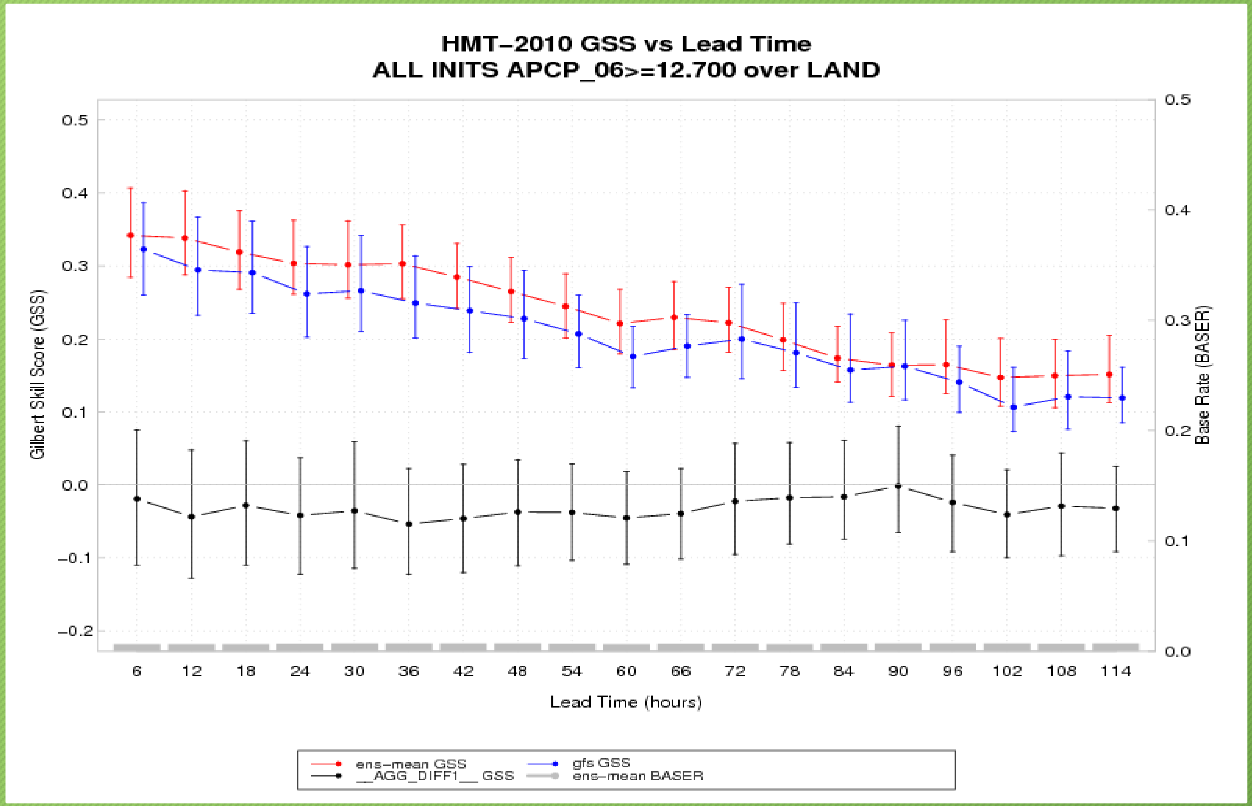

**Period: 20140301-20140520**

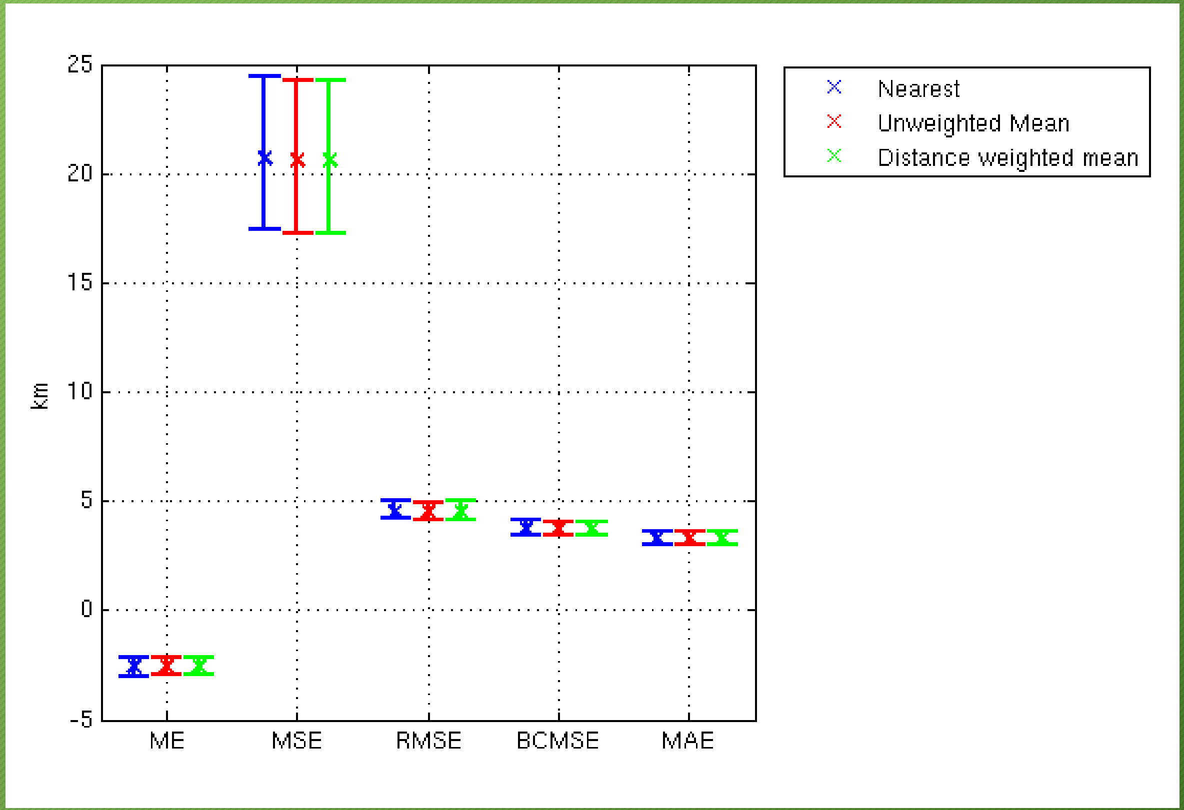
# Confidence Intervals and Interpolation in MET

- Normal Approximation CI
  - Calculated for all statistics for which this is appropriate
- Bootstrapped CI
  - Can be turned on in config file
  - Number of repetitions are user defined

## Interpolation for Point Data

Nearest Neighbor, Unweighted Mean, Distance Weighted Mean, Bilinear Interpolation

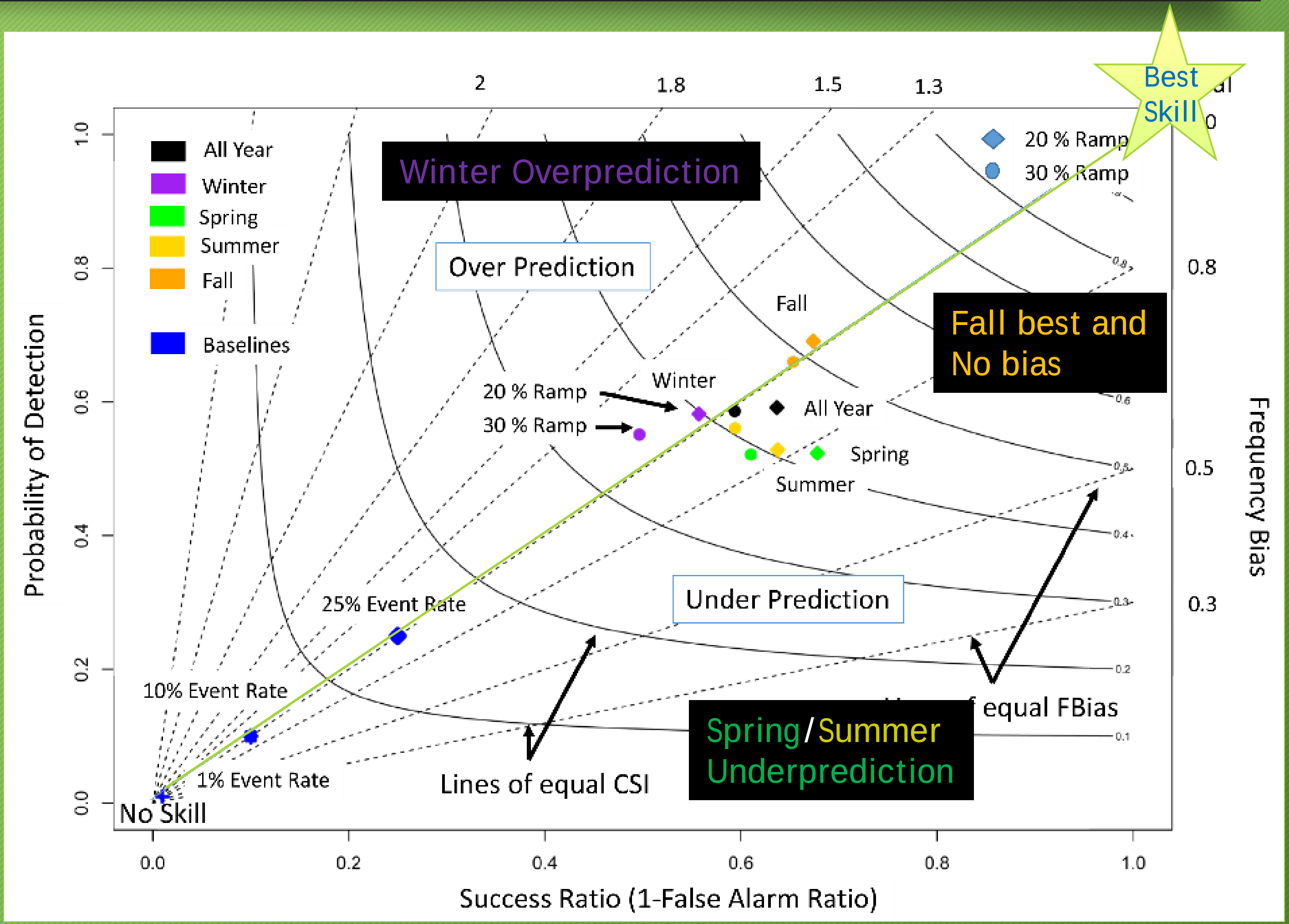


# Performance Diagram for Categorical Statistics

**<u>Performance Diagram for Events</u>**
Allows user to assess skill using multiple metrics & stratify by season, cloud regime, event definition, etc…

*Left: Probability of Detection*
*Bottom: 1-False Alarm Ratio*
*Right: Frequency Bias*
*Curved: Critical Success Index*



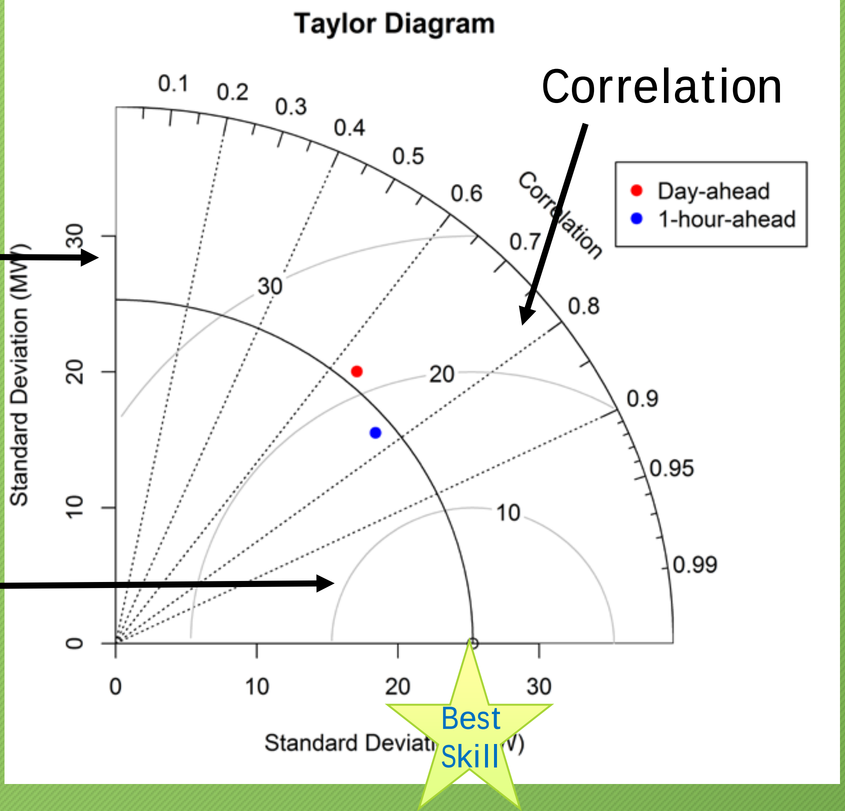

Made using R-statistics

# Taylor Diagram for Continuous Statistics

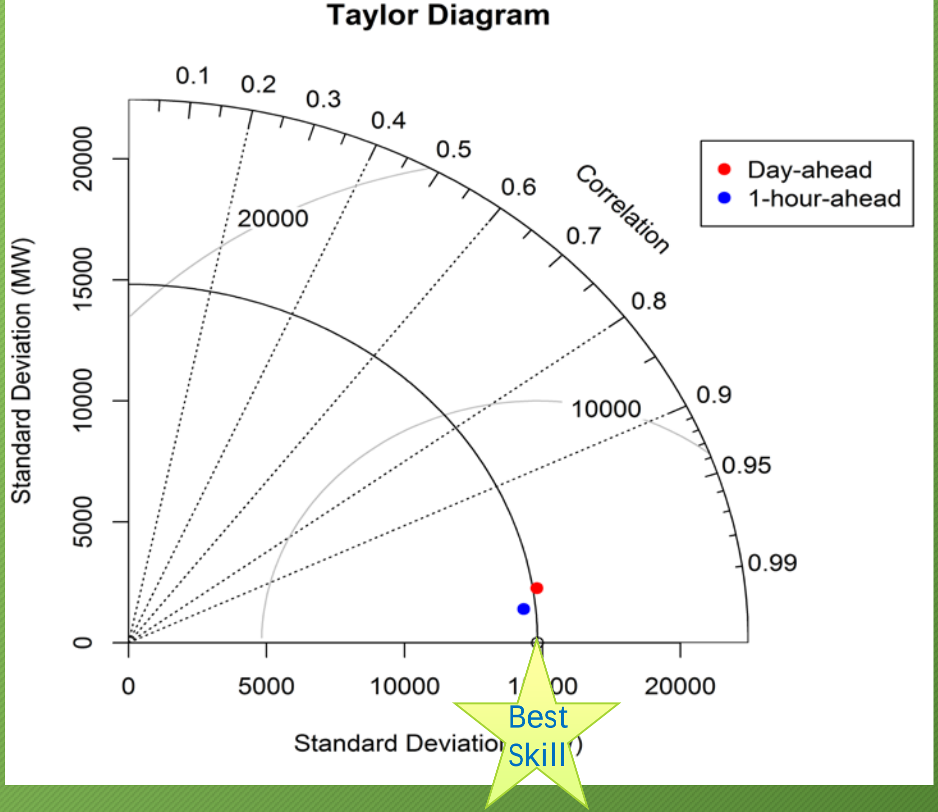
Standard Deviation

RMSE

One Plant




Western Interconnect




Made using R-statistics

Slide Courtesy of Angela Benedetti

# IMPACT OF AEROSOL DIRECT EFFECT ON TC TRACK

## Gordon Track Forecast (10-20 Sep 2006)

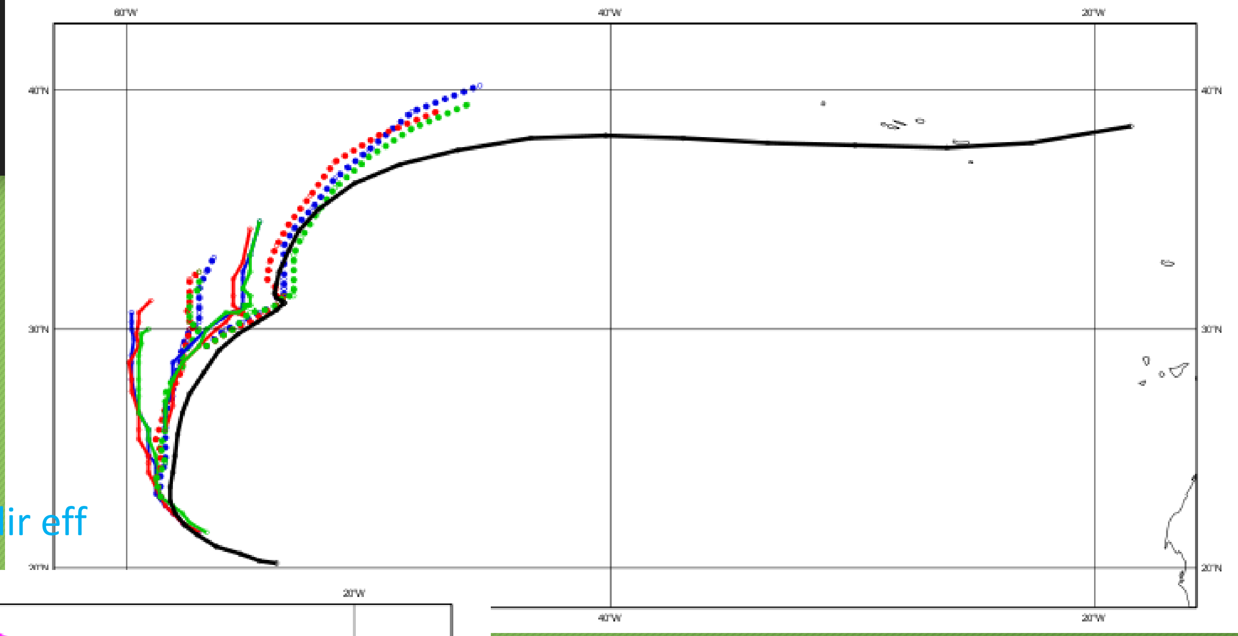
Single wavelength AOD

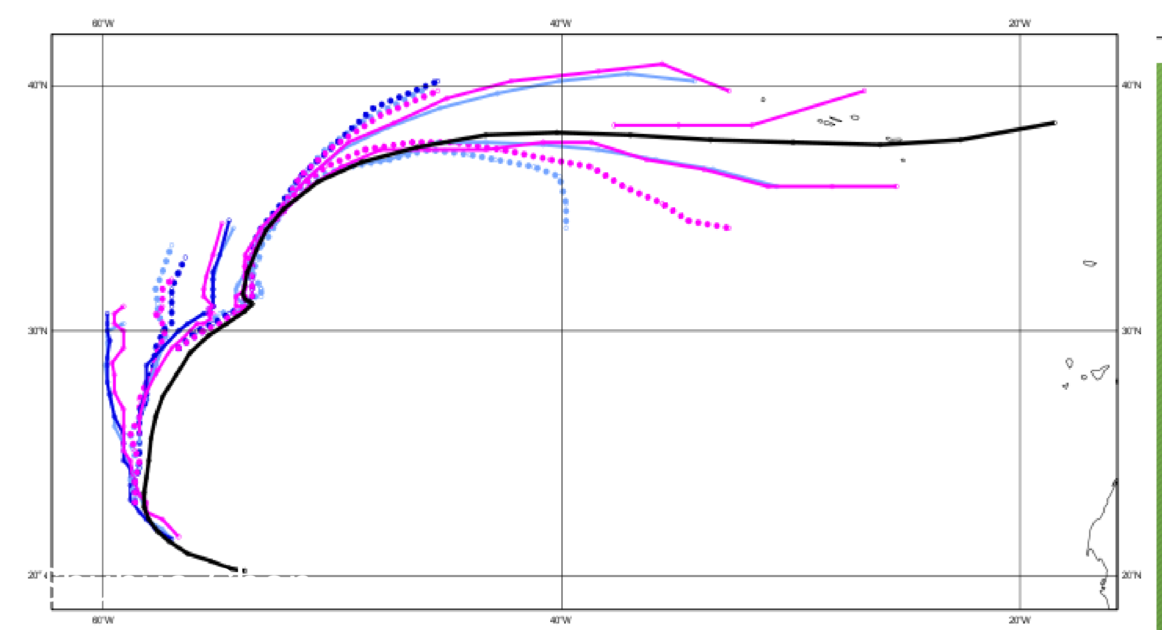
Single wavelength AOD - no direct effect

NO AOD assimilation

Multi-wavelength AOD

Multi-wavelength AOD – no dir eff

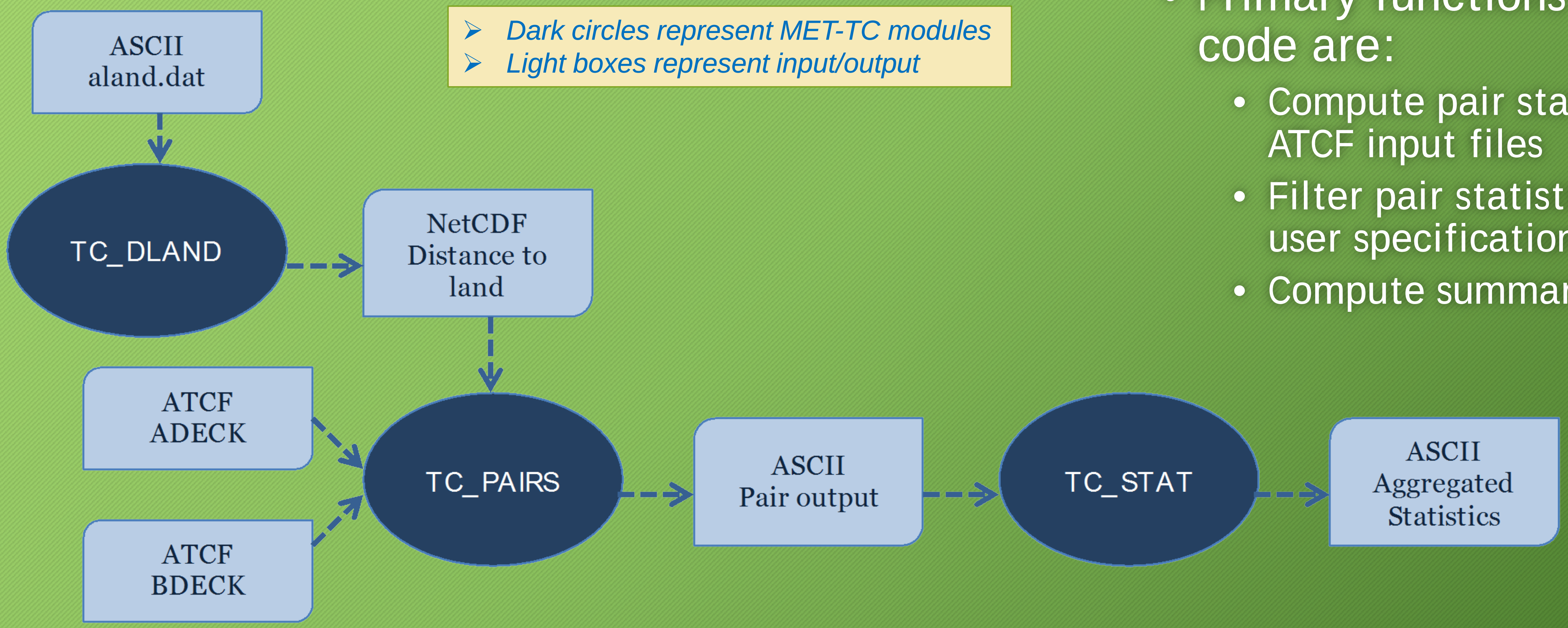



**I would use**

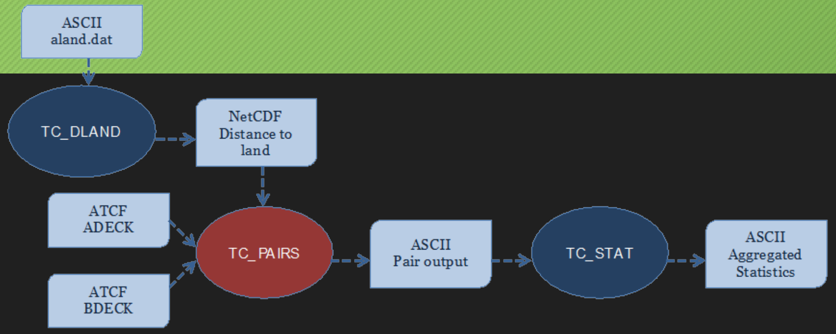
- **My own code to write tracks in ATCF format**
- **MET-TC to evaluate track error**
- **MET-Viewer or R-statistics to calculate pairwise differences to determine statistical significance**

# MET-TC components



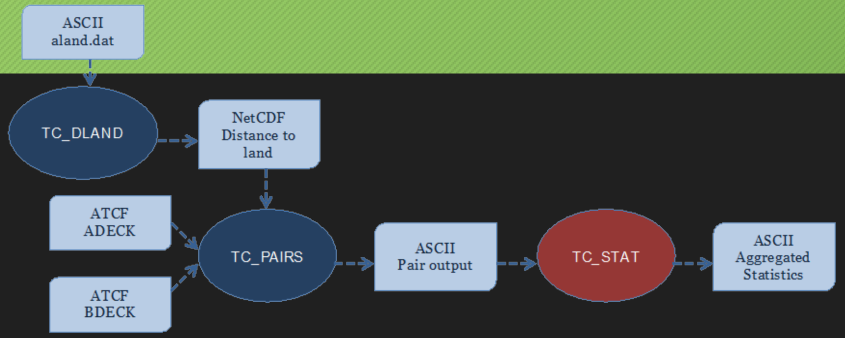

- Primary functions of the code are:
  - Compute pair statistics from ATCF input files
  - Filter pair statistics based on user specifications
  - Compute summary statistics

# TC_PAIRS




- Produces pair statistics on independent model input or user-specified consensus forecasts and "interpolated" forecasts (similar to those discussed in JOHN KNAFF's TALK)
- Matches forecast with reference TC dataset (most commonly Best Track Analysis)
- Pair generation can be subset based on user-defined filtering criteria
- ASCII pair output allows for new or additional analyses to be completed without performing full verification process

# TC_STAT




- Provides summary statistics and filtering jobs on TCST output

✓Filter job

- Based on Thresholding
- Flexible definition of "Rapid Intensification" and "Rapid Weakening" events for diagnostic studies

✓Summary job:

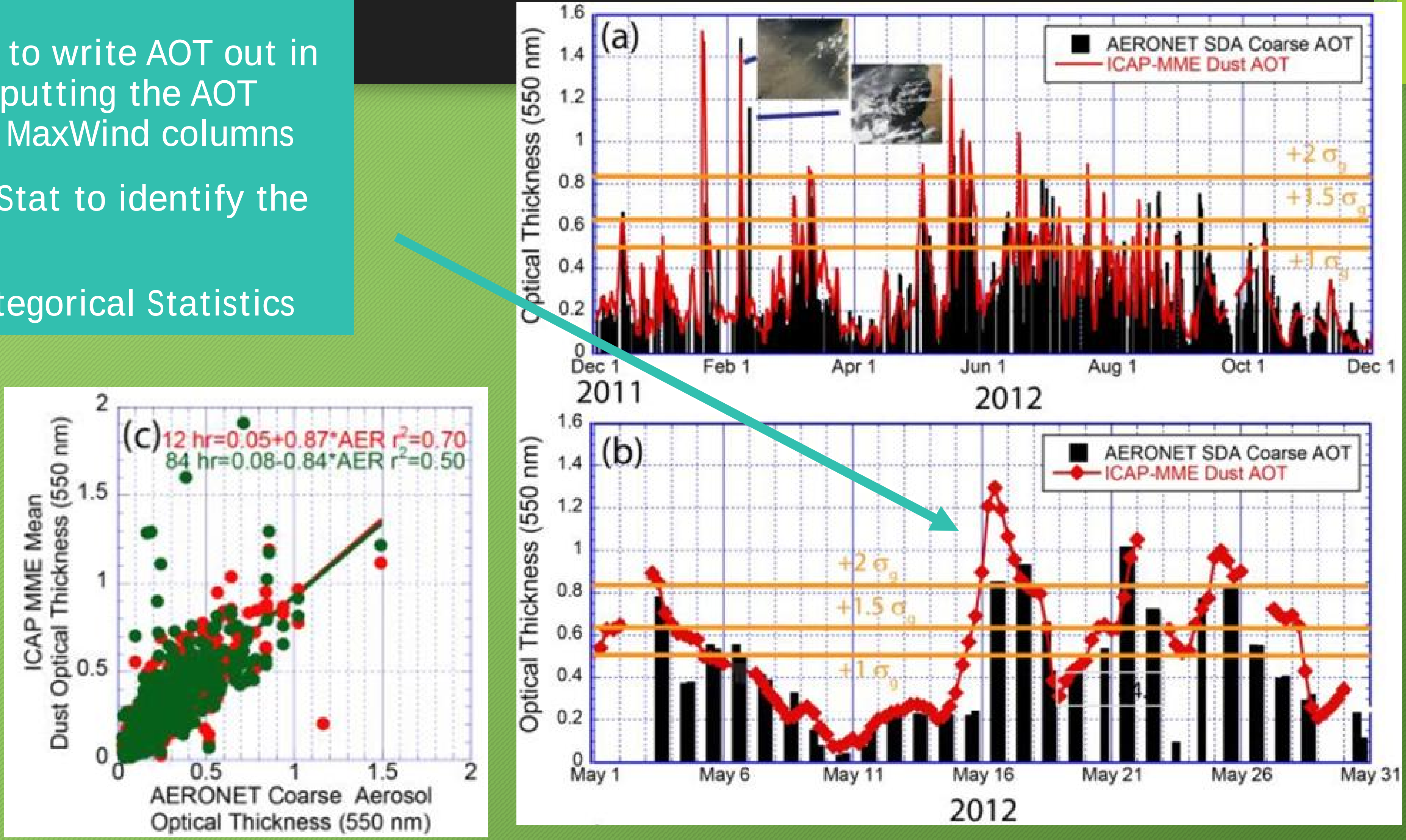
- Produces summary statistics on specific column of interest

# An easy case
# Good News: Cape Verde

Slide courtesy of Jeff Reid

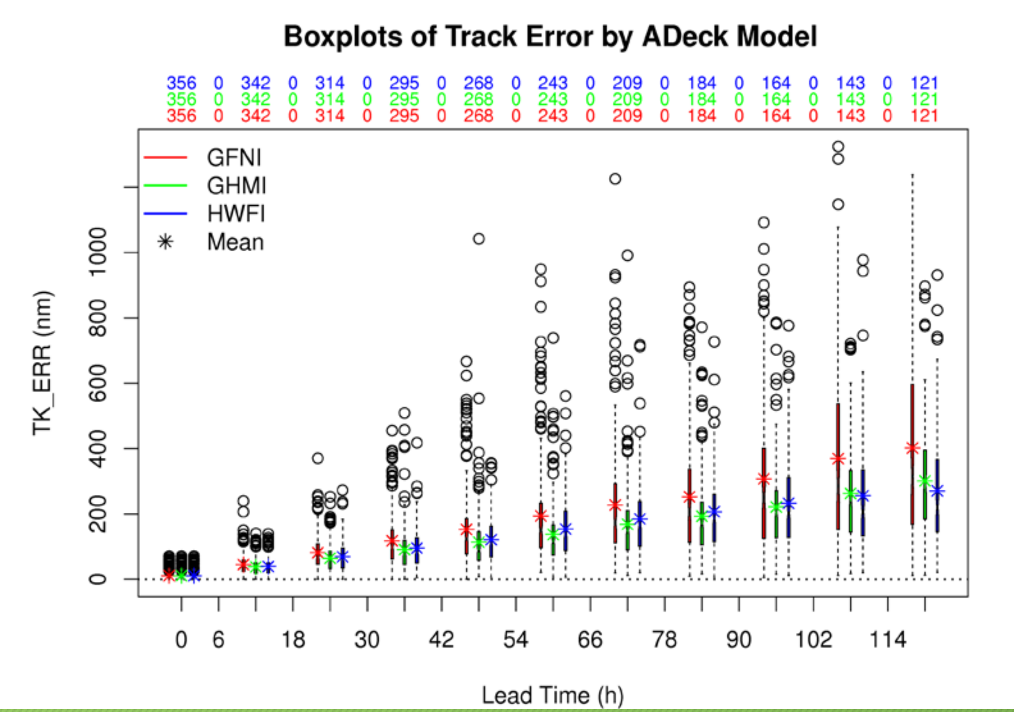
<u>I would use:</u>

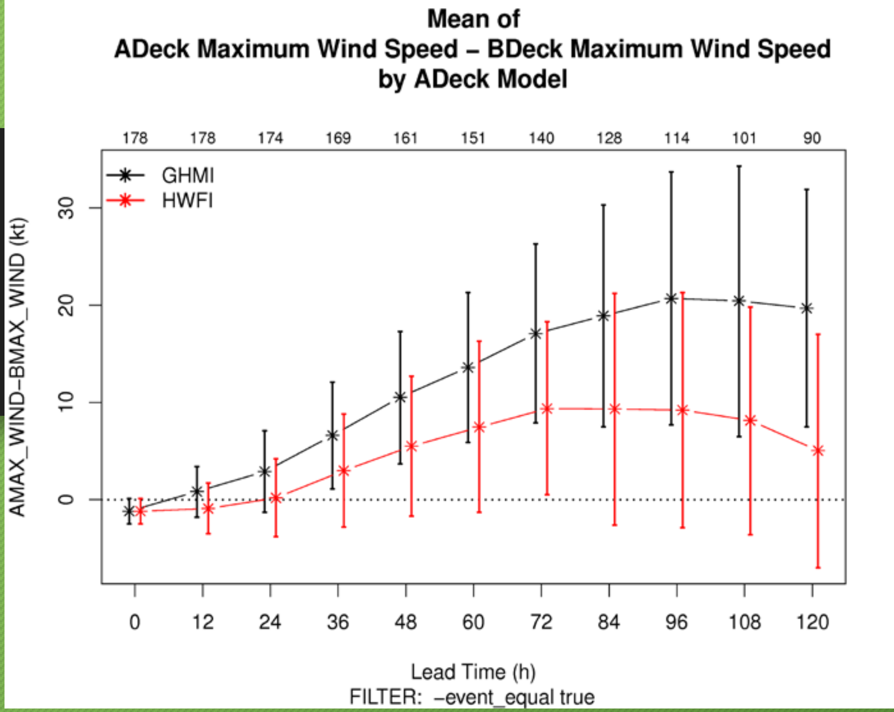
- My own code to write AOT out in ATCF format putting the AOT values in the MaxWind columns
- MET TC – TC-Stat to identify the event
- Calculate Categorical Statistics

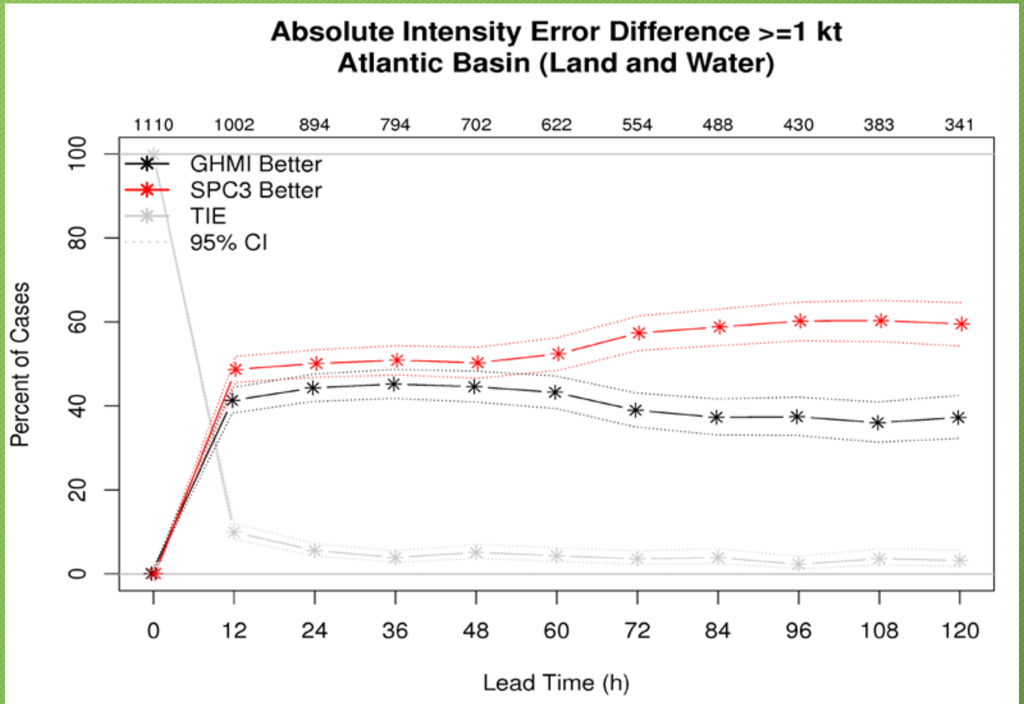


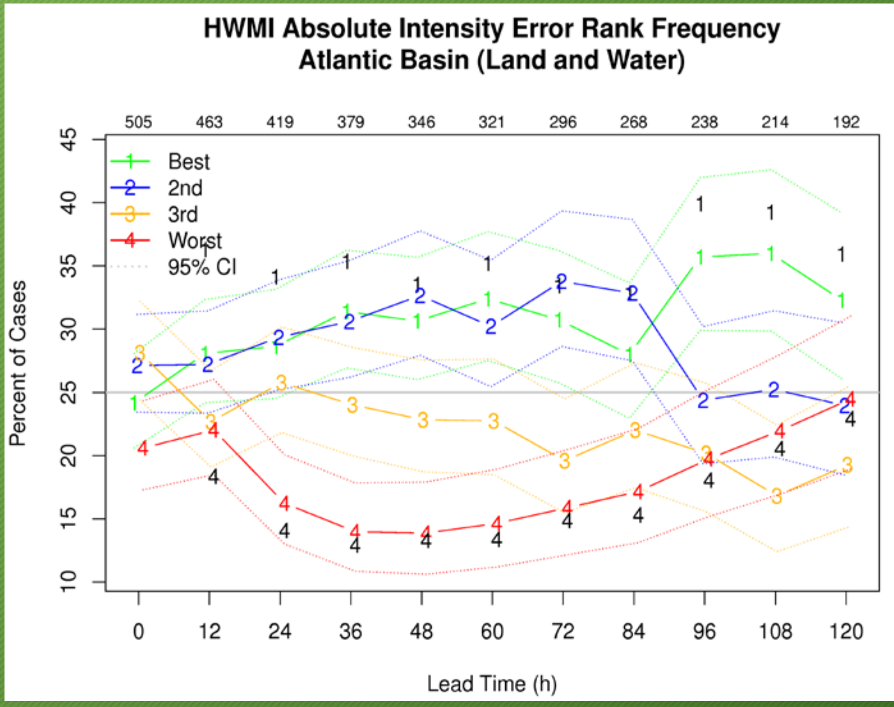
# Graphics tools- examples

R-Statistics scripts available to plot



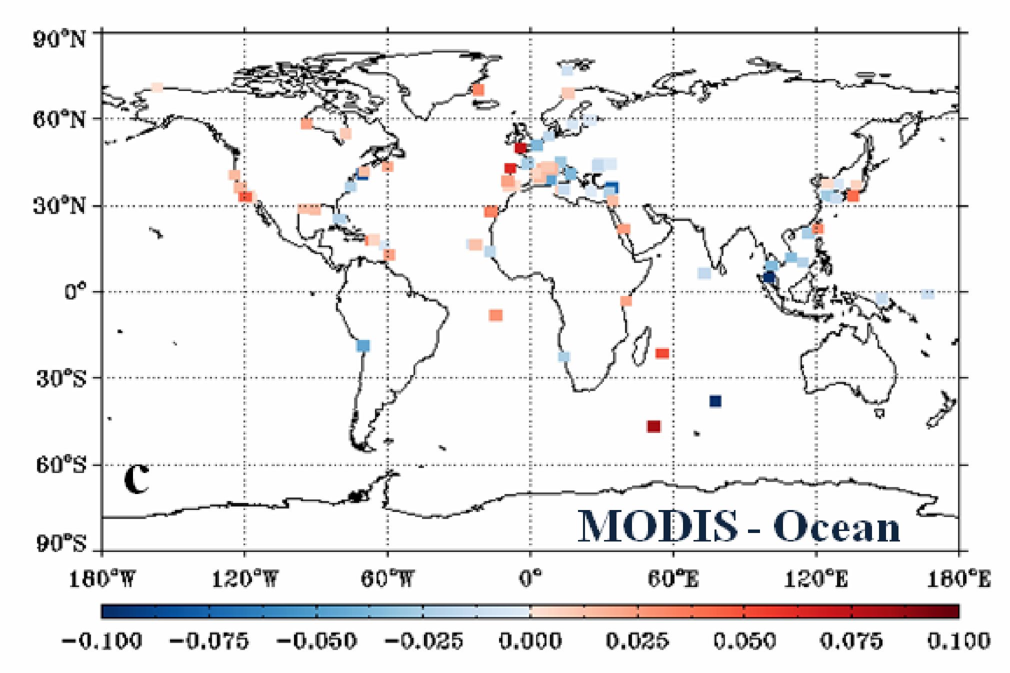



# VIIRS Aerosol Cal/Val

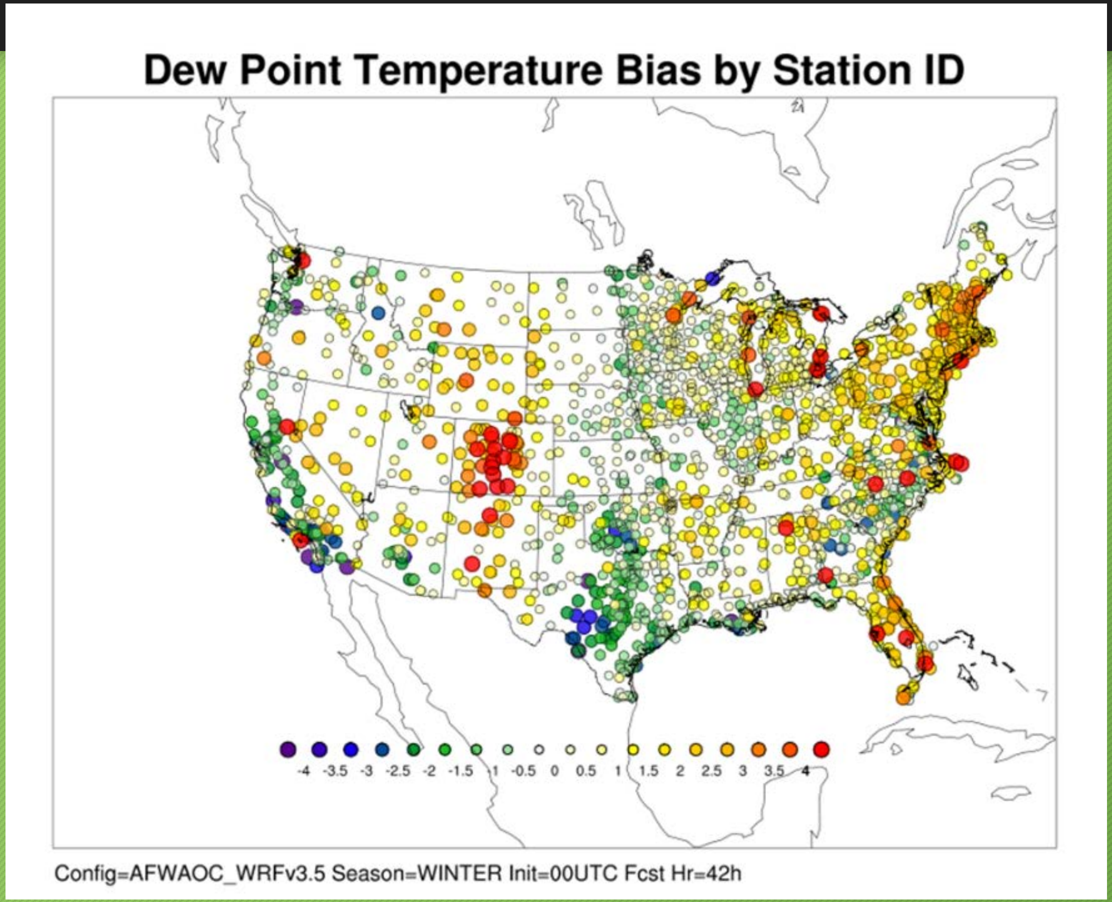





If I wanted a plot like this, I would use:

- Point-stat to calculate the statistic
- Stat-analysis "-by_case" option to "aggregate" the statistic through time for each point
- NCL to plot results

# Geographic Representation of Scores
(Stat-Analysis "by-case" option)




- Accumulates statistics separately for each grid location over a series
  - Time
  - Height
  - Other series
- Accumulate over
  - Stations
  - Grids

# Geographic Representation of Scores

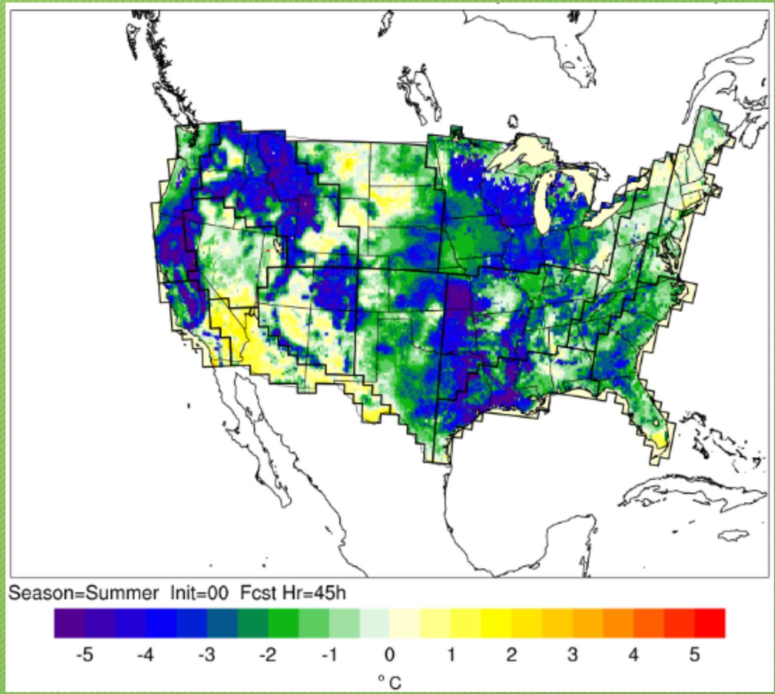
(Series-Analysis tool for gridded data)

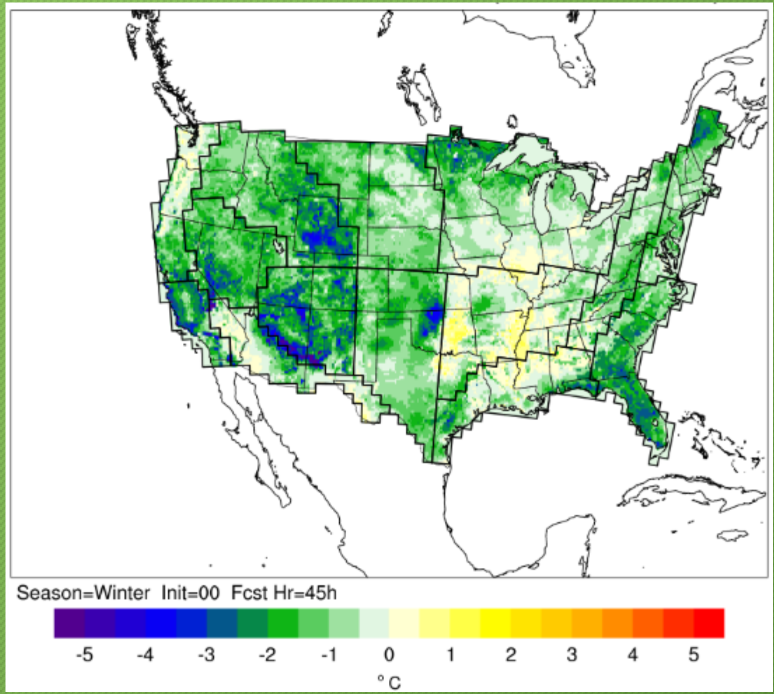
Summer

Winter

**MODEL 1 better**

**MODEL 2 better**




*Soil Temperature (0–10 cm)* – pairwise differences between two model configs

DTC Developmental Testbed Center

Results from DTC Mesoscale Modeling Task

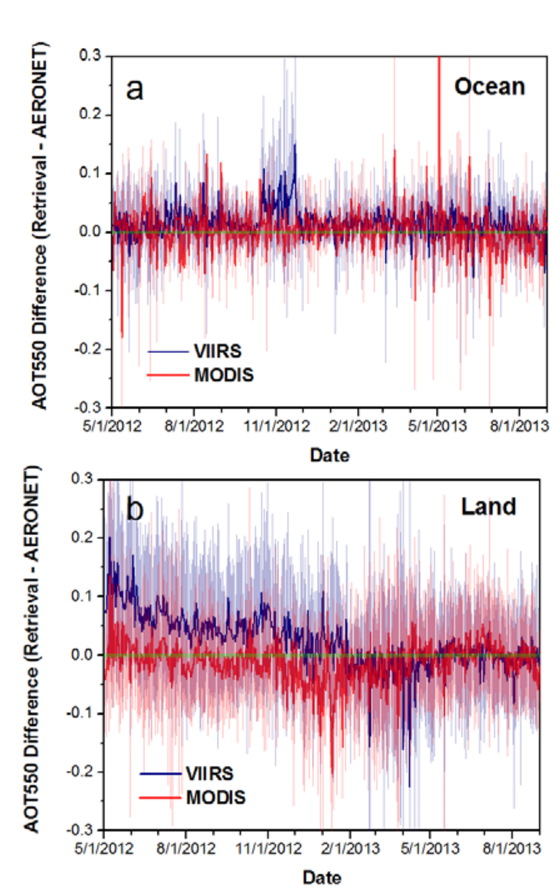
# VIIRS Aerosol Cal/Val

Currently, I would use:

- My own code to right out something called "partial sums" for many cases (not less than 30)
- METViewer (or R-statistics) to calculate the pairwise differences for whatever statistics I'm interested in and plot using bootstrapped confidence intervals
- METViewer (or R-statistics) to plot boxplots to examine distribution of statistics

Eventually, I hope to sometime use:

- My own code to write a generic matched pair file
- Stat-Analysis and calculate statistics directly from matched pairs then calculate the pairwise difference




Hongqing Liu et al., *JGR 2014*

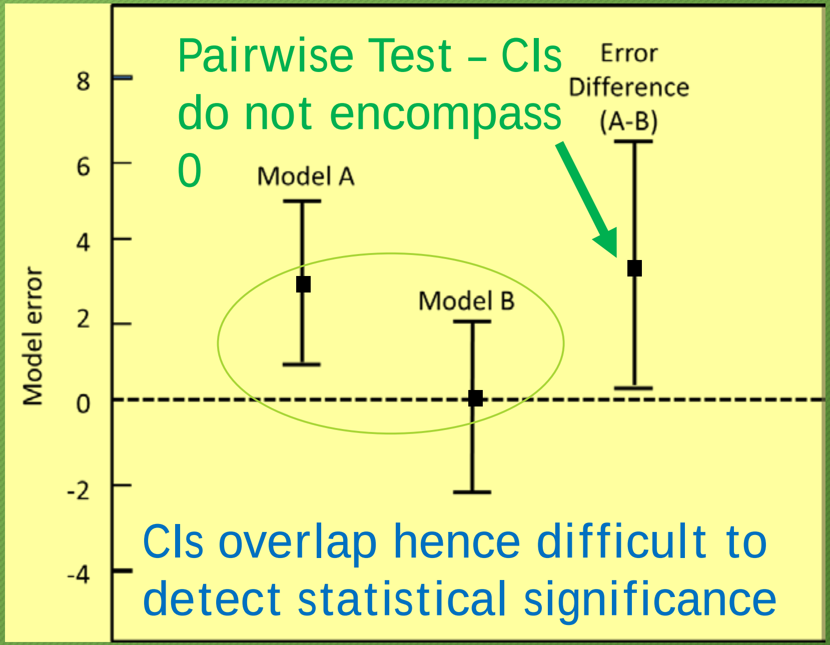
# Paired Tests

| Valid Time | Fcst A | Fcst B | | Fcst A – Fcst B |
|---|---|---|---|---|
| 20140110 11:00 | 4 | - 2 | | = 2 |
| 20140110 12:00 | 1 | - 1 | | = 0 |
| 20140110 13:00 | 2 | - 0 | | = 2 |
| 20140110 14:00 | 5 | - -1 | | = 6 |
| 20140110 15:00 | 3 | - -2 | | = 5 |
| | **3** | **0** | | **3** |

**Mean Error** (arrows point to the Fcst A and Fcst B means)

**Mean Pairwise Error Difference** (arrow points to the Fcst A – Fcst B mean)

**<u>Paired test for model error</u>** shows if the differences in model performance are statistically significant through examination of confidence intervals (CIs)

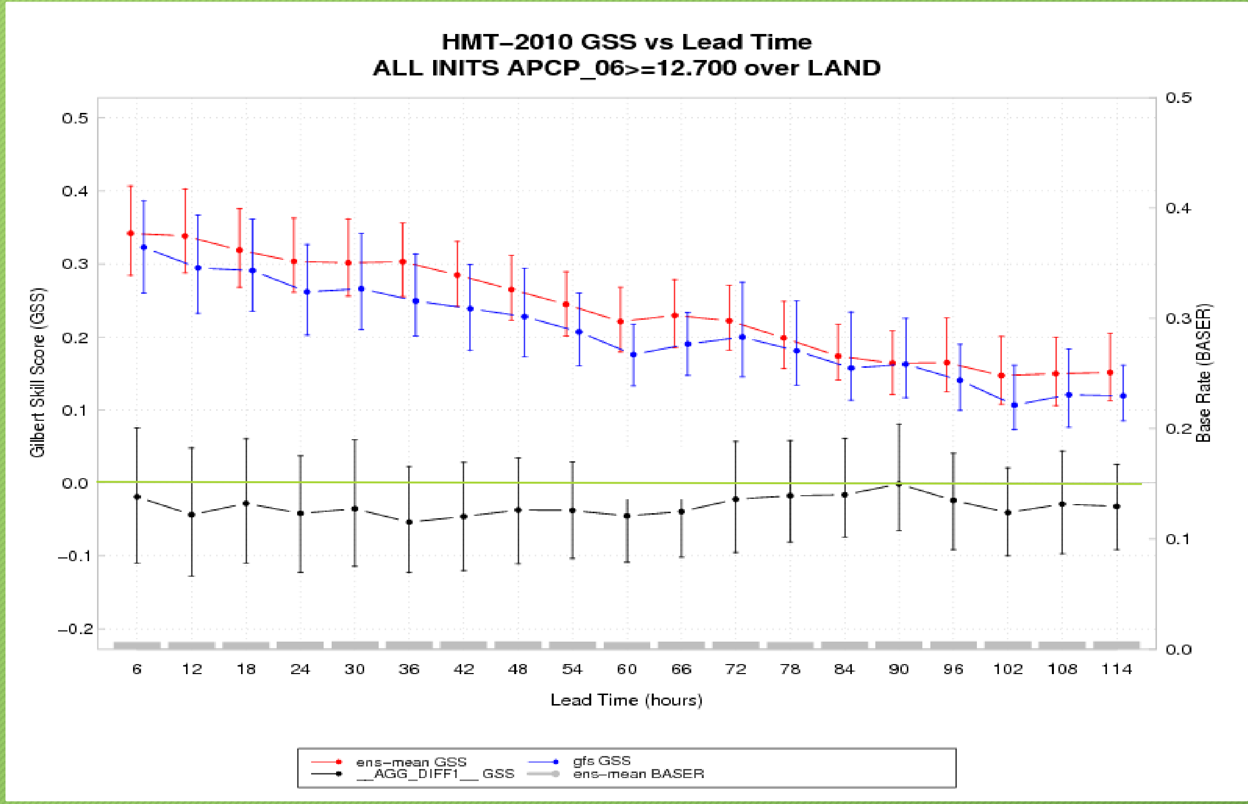


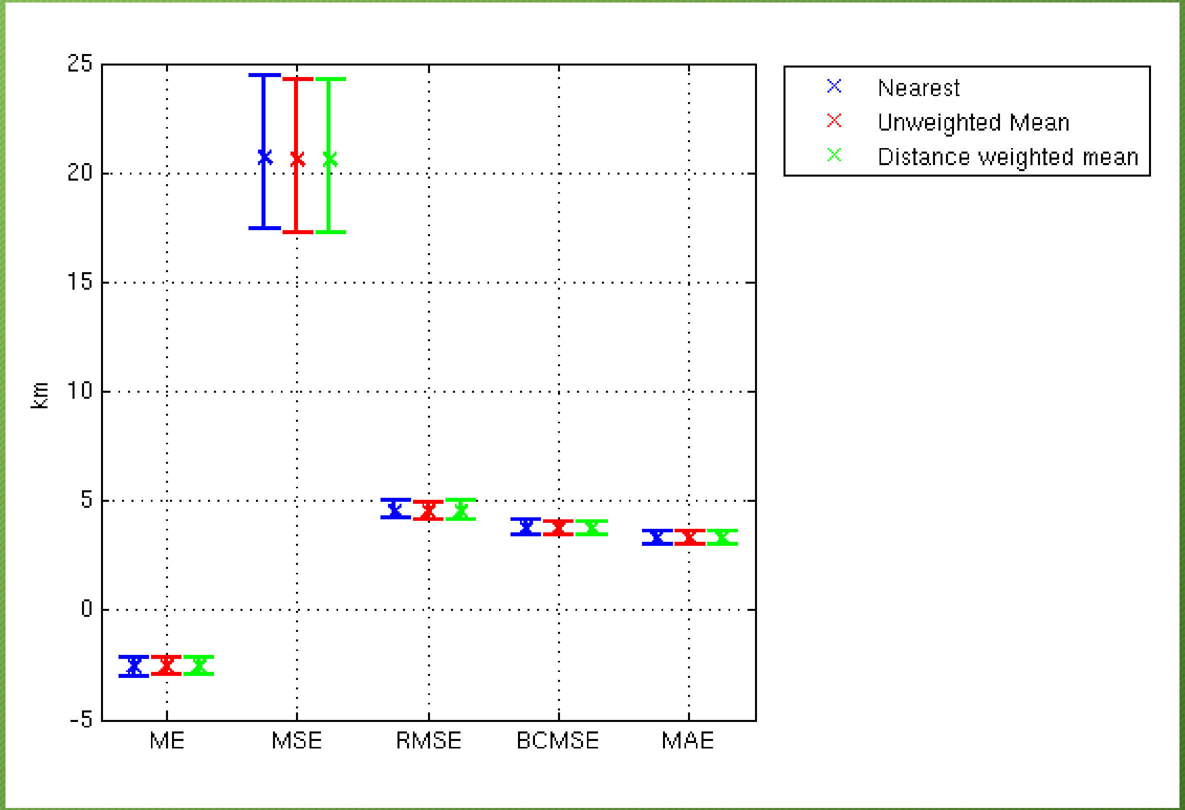
# Confidence Intervals and Interpolation in MET

- Normal Approximation CI
  - Calculated for all statistics for which this is appropriate
- Bootstrapped CI
  - Can be turned on in config file
  - Number of repetitions are user defined

## Interpolation for Point Data

Nearest Neighbor, Unweighted Mean, Distance Weighted Mean, Bilinear Interpolation



# Things you should always do

- Use many cases (minimum 30)
- Use confidence intervals to help determine statistical significance especially when sample size is small
- Look beyond RMSE, MAE, Correlation
- Try to use synthesis tools to look at more than one statistic (i.e. Taylor Diagram, Performance Diagram, Scorecards, etc…)

# Upper Air SS/PS (AFWA – RRTMG)

**SS (light shading)** and **PS (dark shading)** differences for the annual aggregation of upper air temperature and dew point temperature *BCRMSE* and *bias*

| Upper Air Temperature | | Annual | | | | Summer | | | | Winter | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | f12 | f24 | f36 | f48 | f12 | f24 | f36 | f48 | f12 | f24 | f36 | f48 |
| BCRMSE | 850 | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | -- | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG |
| | 700 | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG |
| | 500 | -- | RRTMG | RRTMG | RRTMG | -- | -- | -- | -- | -- | RRTMG | RRTMG | RRTMG |
| | 400 | RRTMG | RRTMG | RRTMG | RRTMG | -- | -- | -- | -- | RRTMG | RRTMG | RRTMG | RRTMG |
| | 300 | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | -- | -- | RRTMG | RRTMG | RRTMG | RRTMG |
| | 200 | -- | -- | AFWA | -- | -- | -- | -- | -- | -- | -- | -- | -- |
| | 150 | -- | -- | -- | -- | -- | AFWA | AFWA | AFWA | -- | -- | -- | -- |
| | 100 | RRTMG | -- | AFWA | -- | -- | -- | -- | -- | RRTMG | -- | -- | -- |
| Bias | 850 | -- | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | -- | RRTMG | RRTMG | RRTMG |
| | 700 | -- | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | RRTMG | -- | RRTMG | RRTMG | RRTMG |
| | 500 | -- | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | -- | -- | RRTMG | RRTMG |
| | 400 | -- | -- | AFWA | AFWA | RRTMG | RRTMG | -- | -- | AFWA | AFWA | AFWA | AFWA |
| | 300 | -- | -- | RRTMG | -- | RRTMG | RRTMG | RRTMG | RRTMG | AFWA | AFWA | AFWA | AFWA |
| | 200 | -- | RRTMG | RRTMG | RRTMG | -- | -- | -- | -- | RRTMG | RRTMG | RRTMG | RRTMG |
| | 150 | AFWA | AFWA | RRTMG | RRTMG | AFWA | AFWA | AFWA | AFWA | -- | -- | -- | -- |
| | 100 | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | -- | -- | -- | -- |

| Upper Air Dew Point Temperature | | Annual | | | | Summer | | | | Winter | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | f12 | f24 | f36 | f48 | f12 | f24 | f36 | f48 | f12 | f24 | f36 | f48 |
| BCRMSE | 850 | RRTMG | RRTMG | RRTMG | RRTMG | -- | -- | RRTMG | RRTMG | RRTMG | -- | RRTMG | -- |
| | 700 | RRTMG | -- | -- | -- | -- | -- | -- | -- | RRTMG | RRTMG | -- | -- |
| | 500 | -- | -- | -- | -- | -- | -- | -- | RRTMG | -- | -- | -- | -- |
| Bias | 850 | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | RRTMG | -- | AFWA | AFWA | AFWA |
| | 700 | AFWA | AFWA | AFWA | AFWA | RRTMG | RRTMG | RRTMG | RRTMG | AFWA | AFWA | AFWA | AFWA |
| | 500 | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA | AFWA |

**Statistical Significance (light shading)**

- **Differences pass the test**

**Practical Significance (dark shading)**

- **Which SS differences are greater than the observation uncertainty**

DTC
Mesoscale Modeling Task
Developmental Testbed Center

# Thank Yous and Further Information

*DTC would like to thank you for your interest and the assistance of all of our collaborators…*

JNT: http://www.ral.ucar.edu/jnt
DTC: http://www.dtcenter.org
MET: http://www.dtcenter.org/met/users

Email: jensen@ucar.edu
met-help@ucar.edu

Support for the Developmental Testbed Center (DTC), is provided by NOAA, AFWA NCAR and NSF