The Role in Verification in R2O Testing and Evaluation

Tara L. Jensen[1], and M. P. Mittermaier[2], T. Fowler[1], B. G. Brown[1], J. Halley Gotway[1], R. Bullock[1], J. K. Wolff[1], M. Harrold[1], E. Tollerud[3], L. Nance[1], and Y. H. Kuo[1]

1. NCAR/RAL, Developmental Testbed Center (DTC)
2. UK Met Office
3. CIRA, NOAA/ESRL, Developmental Testbed Center (DTC)

95th AMS – Fifth Conference on Research to Operations Transiztion
4-8 February 2015 Phoenix, AZ

Developmental Testbed Center

NCAR

Why verify?

- Purposes of verification (traditional definition)

- Administrative purpose
  - Monitoring performance
  - Choice of model or model configuration (has the model improved?)

- Scientific purpose
  - Identifying and correcting model flaws
  - Improved forecasts

- Economic purpose
  - Improved decision making
  - “Feeding” decision models or decision support systems

Identifying verification goals

What *questions* do we want to answer?

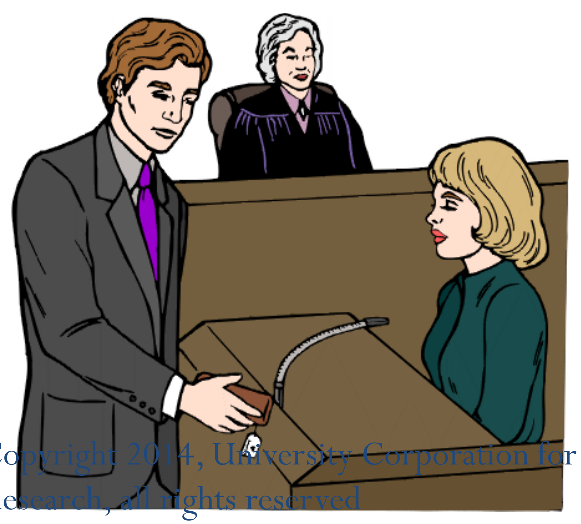
- Examples:
  - ✓In what locations does the model have the best performance?
  - ✓Are there regimes in which the forecasts are better or worse?
  - ✓Is the probability forecast well calibrated (i.e., reliable)?
  - ✓Do the forecasts correctly capture the natural variability of the weather?

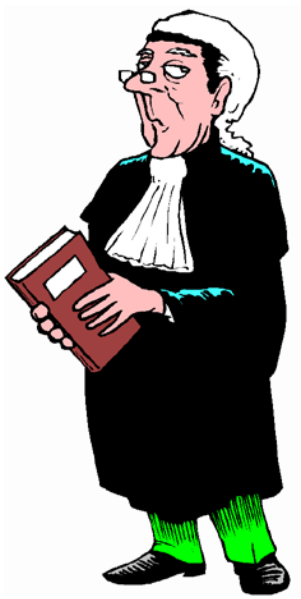
Basic guide for developing verification studies

Identify multiple ***verification attributes*** that can provide answers to the questions of interest

Select ***measures and graphics*** that appropriately measure and represent the attributes of interest

Identify a ***standard of comparison*** that provides a reference level of skill (e.g., persistence, climatology, old model)





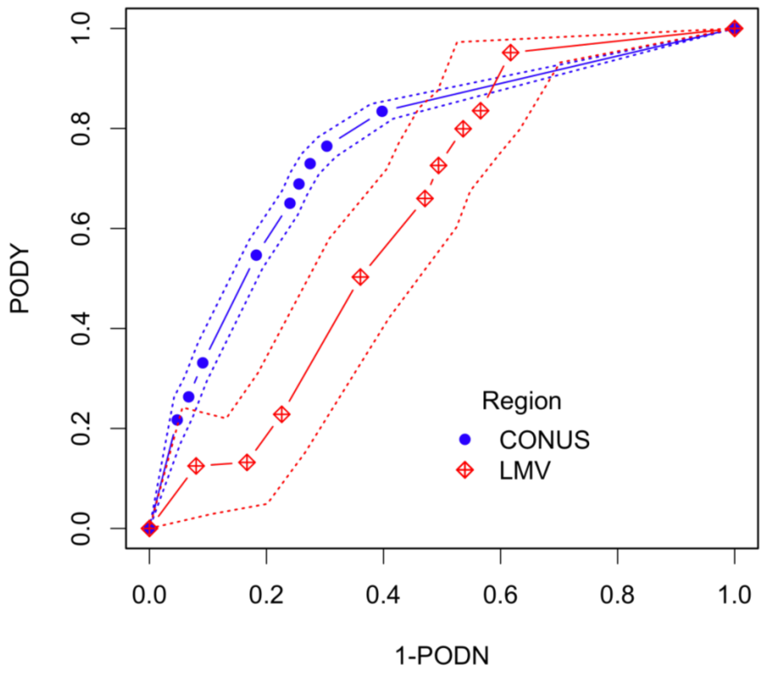
Confidence Intervals

Uncertainty in scores and measures should be estimated whenever possible!

- Uncertainty arises from
  - Sampling variability
  - Observation error
  - Representativeness differences

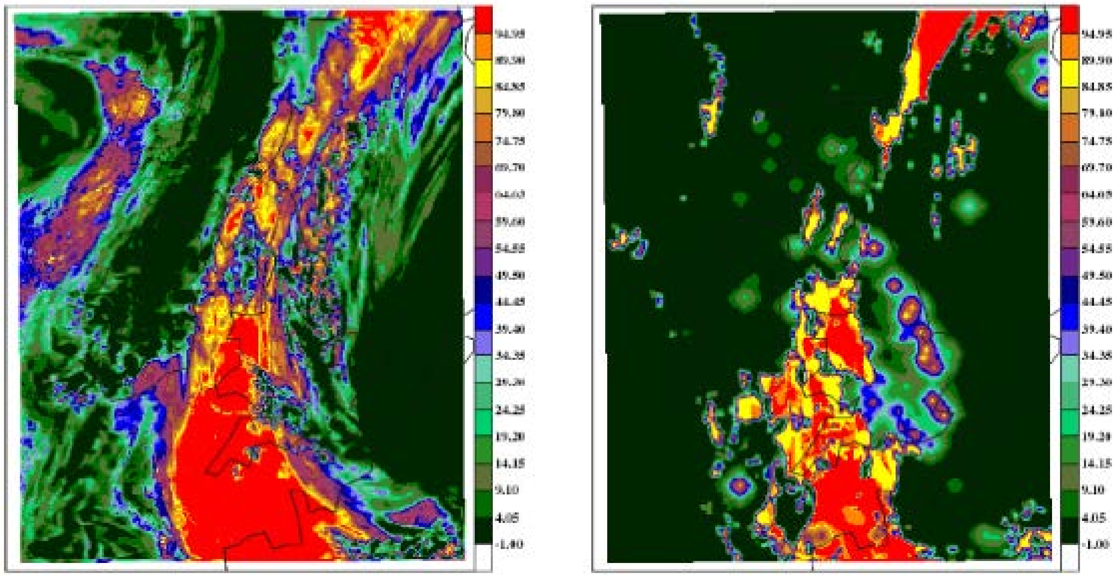
- Erroneous conclusions can be drawn regarding improvements in forecasting systems and models without CIs

- Methods for *confidence intervals* and *hypothesis tests*
  - Parametric (i.e., depending on a statistical model)
  - Non-parametric (e.g., derived from re-sampling procedures, often called "bootstrapping")

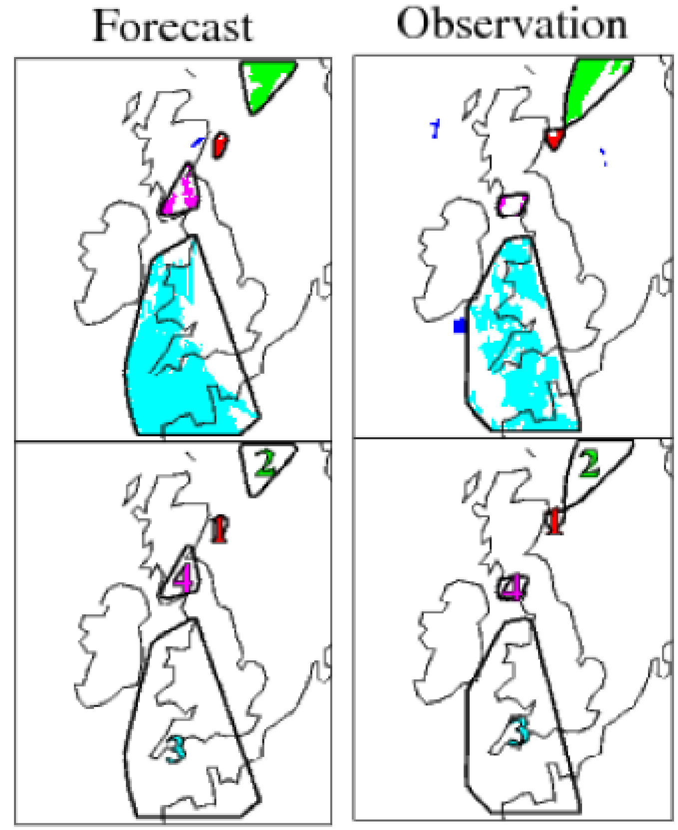


UK4 t+12h @ 1500 UTC

Analysis @ 1500 UTC

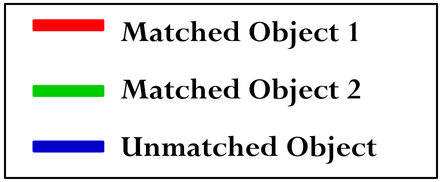


Courtesy of Marion Mittermaier
The Met Office

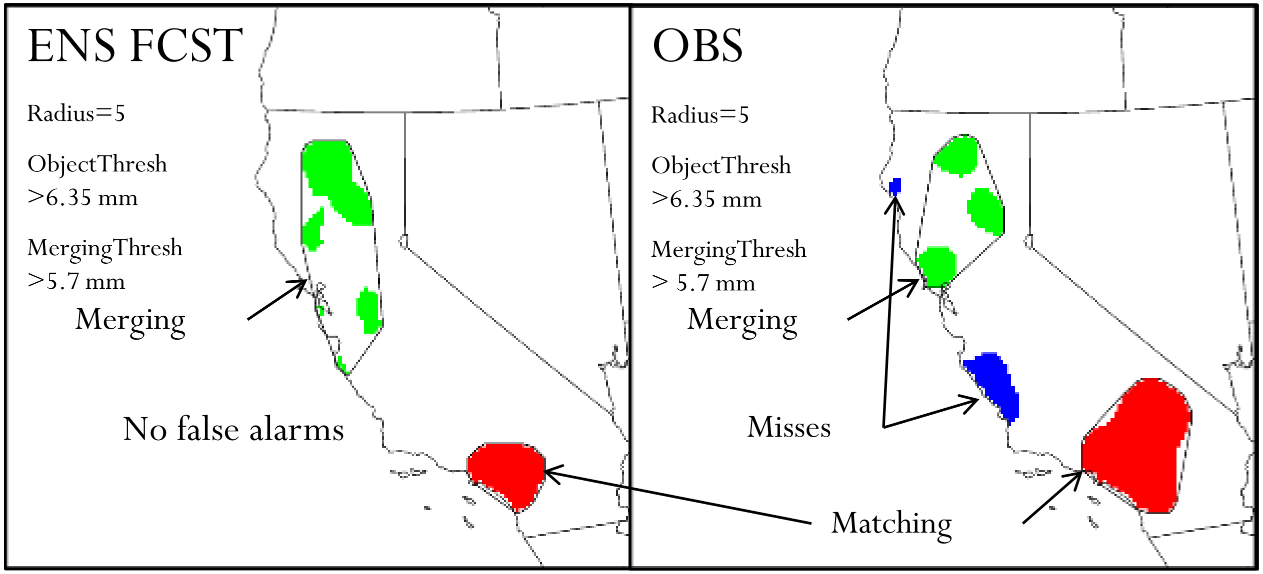


Spatial Methods

Method for Object-based Diagnostic Evaluation (MODE)



How it works



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

This:	Instead of this:
30% Too Big (area ratio=1.3)	POD = 0.35
Shifted west 1 km (centroid distance = 1km)	FAR = 0.7235
Rotated 15° (angle diff = 15%)	CSI = 0.1587
Peak Rain 1/2" too much (diff in 90th percentile of intensities = 0.5)	

Example of Advanced Techniques in R2O

Met Office

Temporal evolution

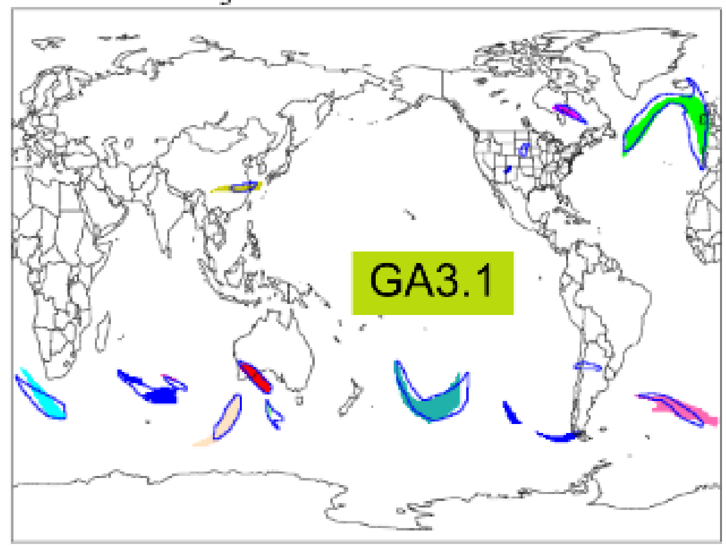
- Older N320 trial
  250 hPa winds > 60 m/s
  at forecast lead time of t+96h
  from the 12Z initialisation
  compared to EC analyses

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

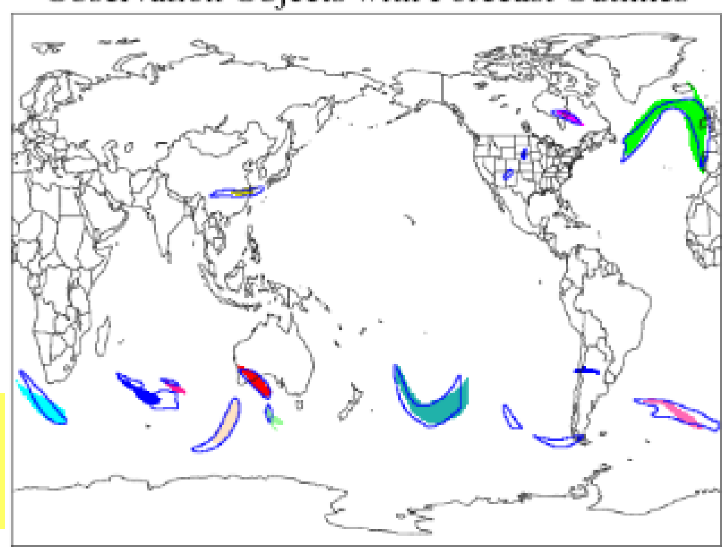
Slide courtesy of Marion Mittermaier
The Met Office – August 2014

Model Evaluations Tool (MET) used to identify objects and synthesize attributes

Forecast Objects with Observation Outlines



Observation Objects with Forecast Outlines



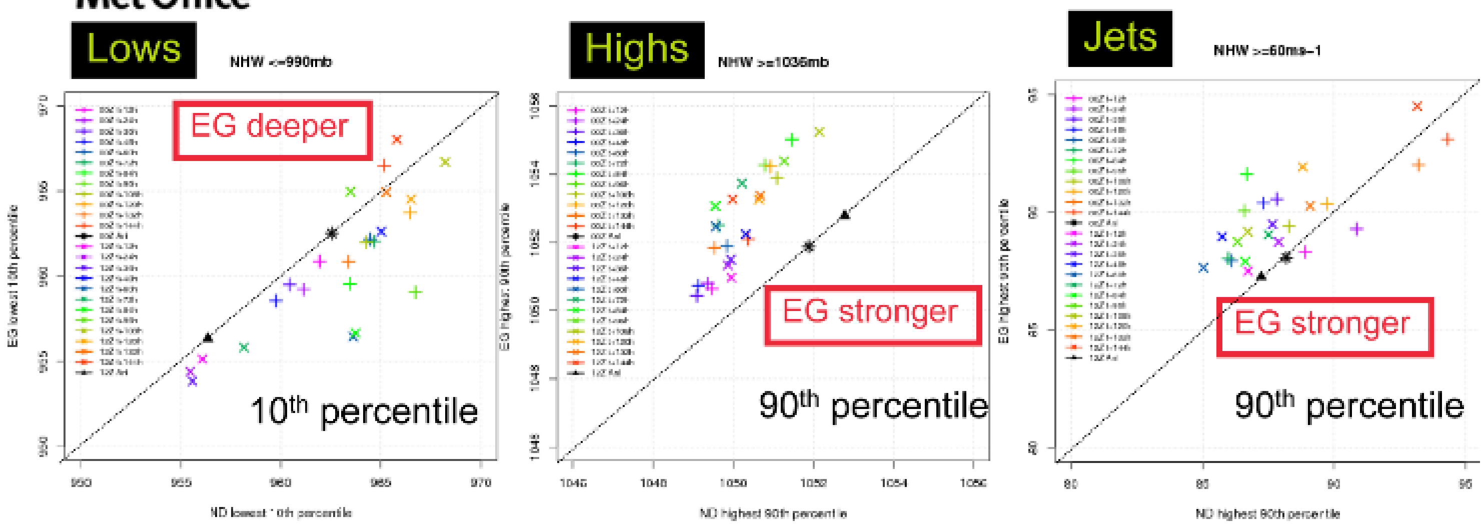
DTC
Developmental Testbed Center

Example of Advanced Techniques in R2O

EC analyses
N768 EG v N512 ND

Met Office

Object intensities

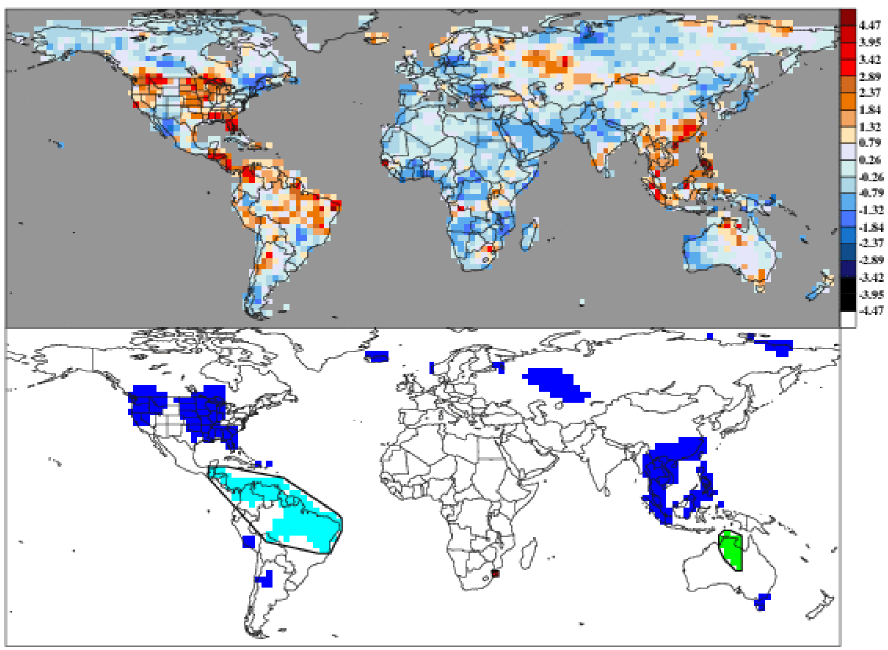


- Do not look at absolute min/max values in objects. Use the 10th or 90th percentile as a more reliable estimate of how the intensity distribution has shifted/changed.
- Lows are deeper, highs and jets are stronger → sharper gradients and a more active energetic model.
- Differences in the 00Z and 12Z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

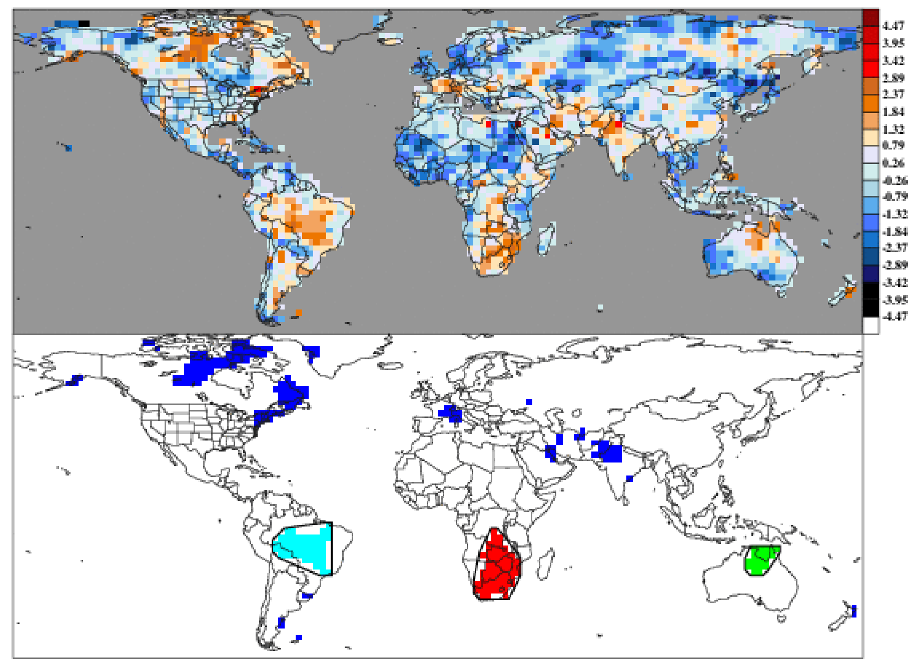
Slide courtesy of Marion Mittermaier
The Met Office – August 2014

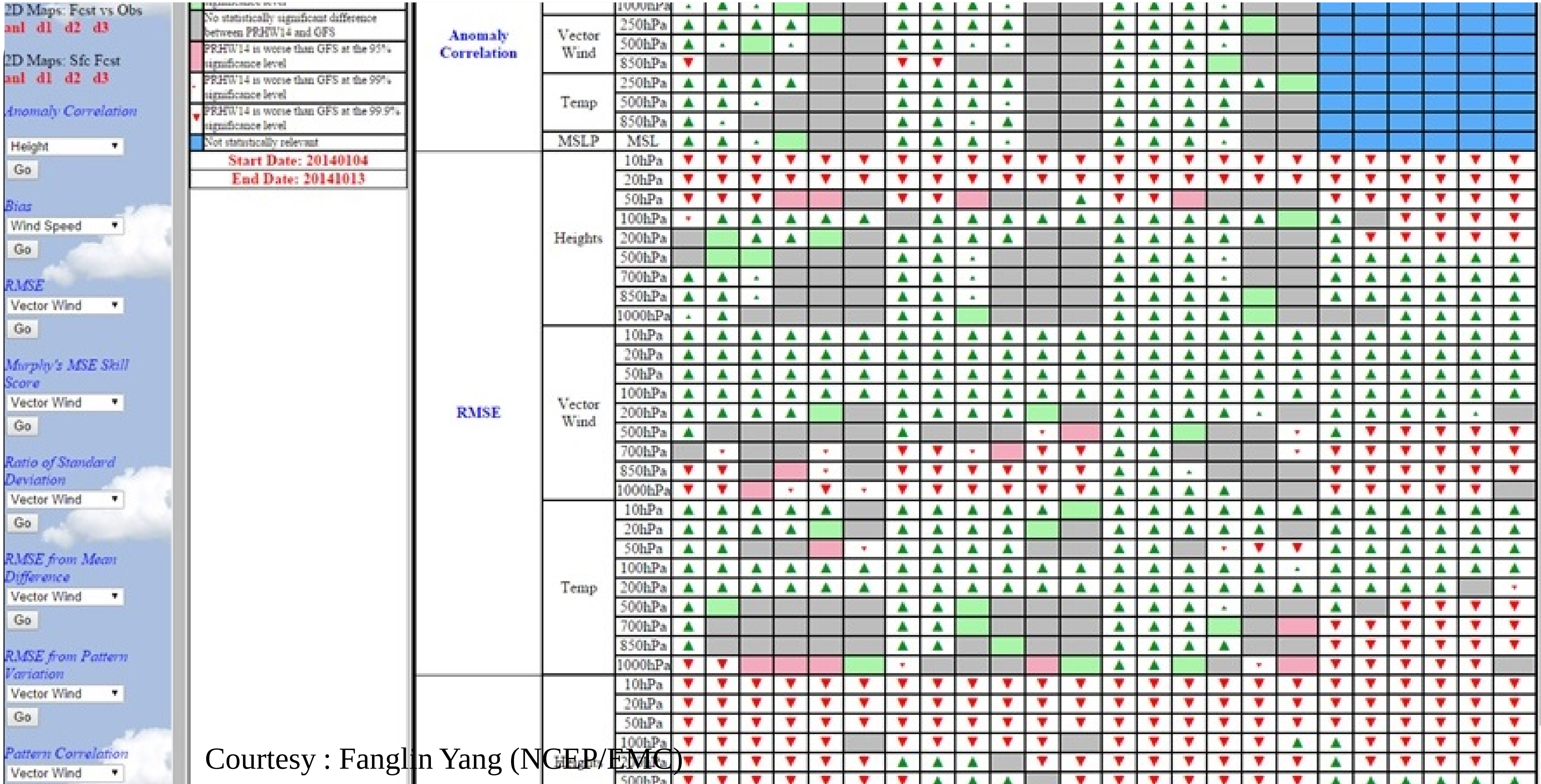
Use of MODE on Climate…

MODE: pdsisc at SFC vs pdsisc at SFC
Forecast



MODE: pdsisc at SFC vs pdsisc at SFC
Observation





Courtesy : Fanglin Yang (NCEP/EMC)

Scorecarding and NWP Indexes

AFWA Configuration Testing | DTC

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	--
BCRMSE	700	RRTMG	--	--	--	--	--	--	--	RRTMG	RRTMG	--	--
BCRMSE	500	--	--	--	--	--	--	--	RRTMG	--	--	--	--
Bias	850	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	RRTMG	--	AFWA	AFWA	AFWA
Bias	700	AFWA	AFWA	AFWA	AFWA	RRTMG	RRTMG	RRTMG	RRTMG	AFWA	AFWA	AFWA	AFWA
Bias	500	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA	AFWA

Model Evaluations Tool (MET) used to calculate the statistics and scripting used to formulate scorecard

Statistical Significance (light shading)

- Differences pass the test

Practical Significance (dark shading)

- Which SS differences are greater than the observation uncertainty

NWP Index

Example – AFWA GO Index – a weighted average of the RMSE values for wind speed, dewpoint temperature, temperature, height, and pressure at several levels in the atmosphere.

Table 10-1. Variables, levels, and weights used to compute the GO Index.

Variable	Level	Weights by lead time			
		12 h	24 h	36 h	48 h
Wind speed	250 hPa	4	3	2	1
	400 hPa	4	3	2	1
	850 hPa	4	3	2	1
	Surface	8	6	4	2
Dewpoint temperature	400 hPa	8	6	4	2
	700 hPa	8	6	4	2
	850 hPa	8	6	4	2
	Surface	8	6	4	2
Temperature	400 hPa	4	3	2	1
	Surface	8	6	4	2
Height	400 hPa	4	3	2	1
Pressure	Mean sea level	8	6	4	2

Model Evaluations Tool (MET) has this capability to compute GO Index but is flexible enough to compute other user-defined NWP Indexes

MET

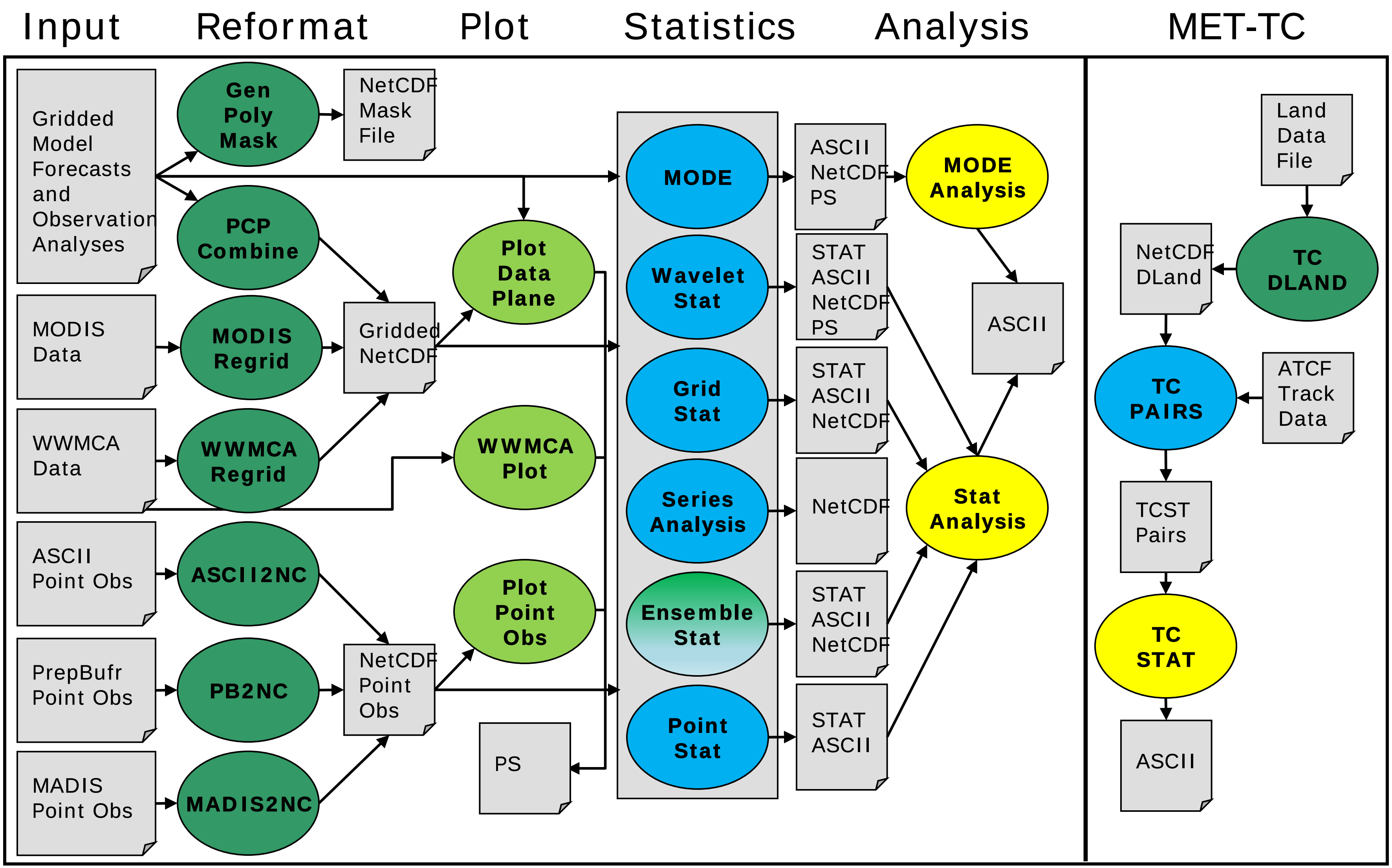
Model Evaluation Tools

Provides a standard set of tools to facilitate reliable, consistent verification across institutions

MET Package

- MET is community code supported by DTC that is free to download (registration required)
  - Approximately 2550 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 in Boulder, CO
  It's not too late to register! Contact me for how to do so.

MET v5.0 Tools (Release: August 2014)



Details on Categorical and Continuous Statistics

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

New in MET v5.0

24 Statistics

25 Statistics

RUN:

POINT-STAT
OR
GRID-STAT

Neighborhood and Ensemble/Probability Statistics

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

RUN: GRID-STAT

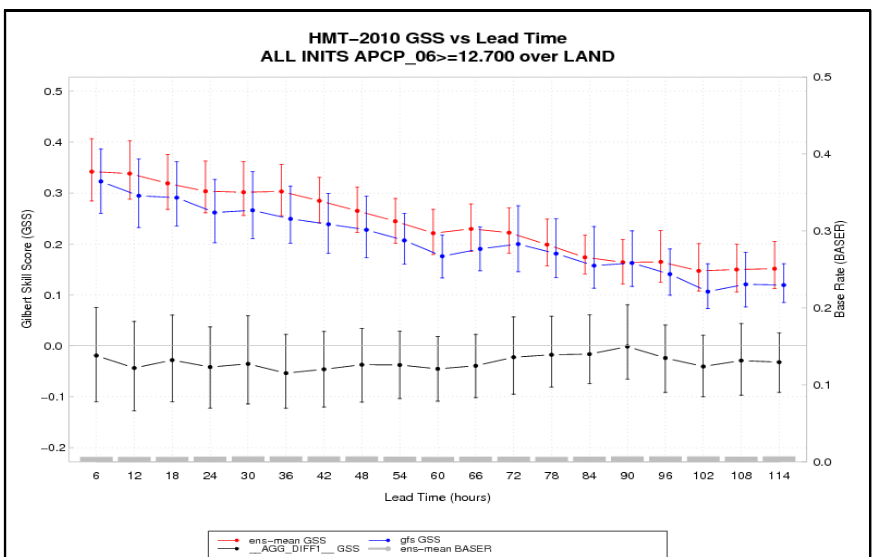
RUN: WAVELET-STAT

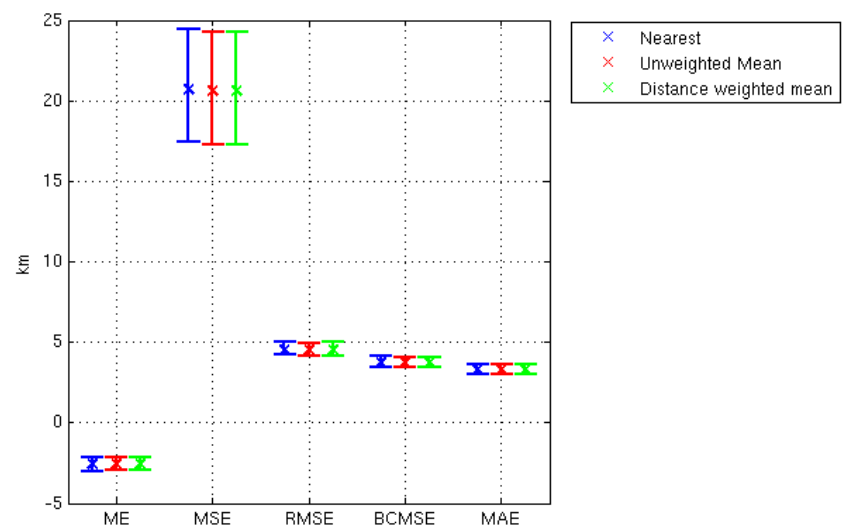
RUN: ENSEMBLE STAT

RUN: POINT-STAT OR GRID-STAT

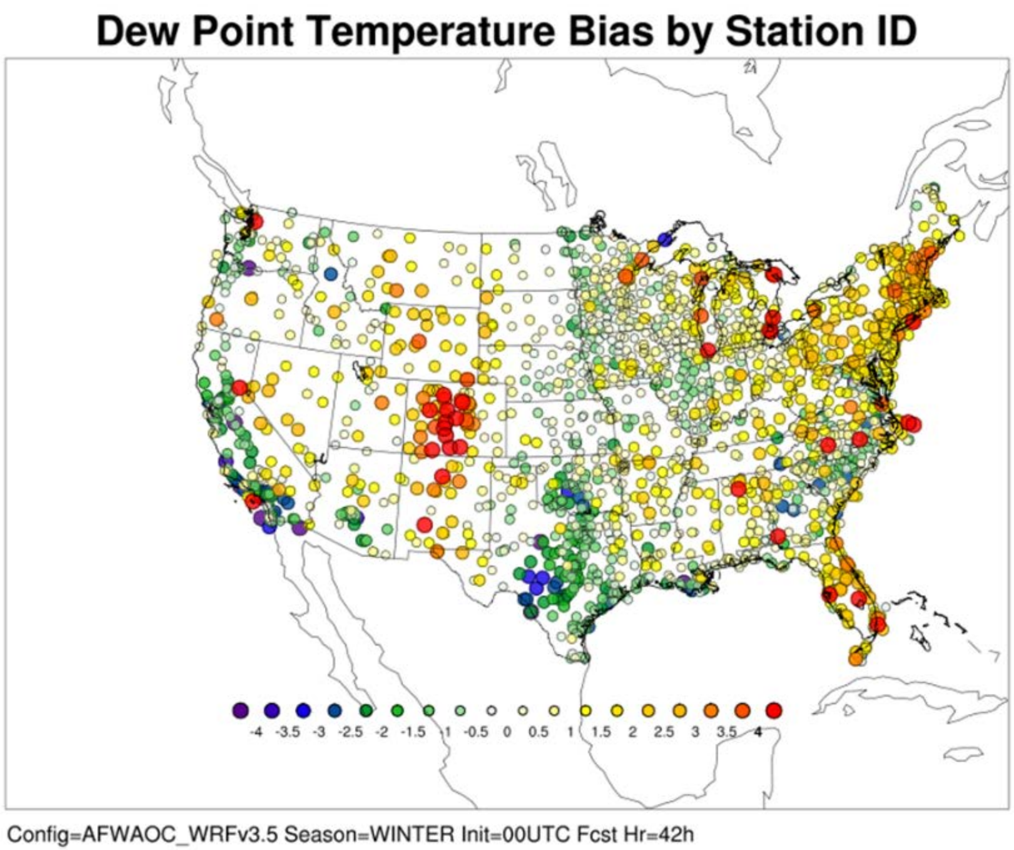
Reason to use MET: Easy use of Confidence Intervals

- 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





Series Analysis tools for geographic representation of scores



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

MET
Model Evaluation Tools

Support for MET has been provided by

AFWA, NOAA and NCAR

through the Developmental Testbed Center (DTC)

Thank You and Further Information

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