HWT-DTC Objective Evaluation 2010 SE Overview

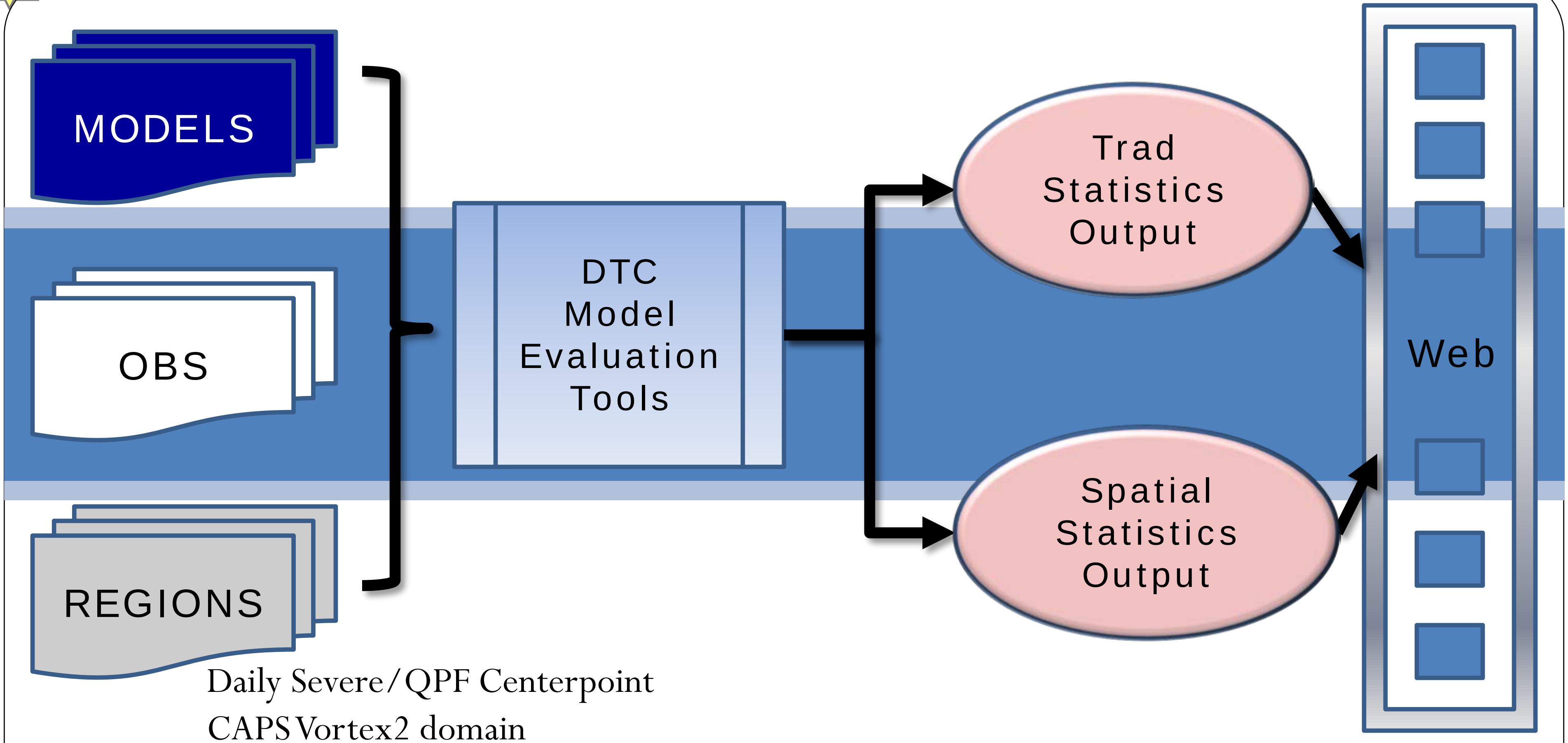
Tara Jensen[1], Dave Ahijevych[1], Michelle Harrold[1], Jamie Wolff[1] , Isidora Jankov[2,3], and Barb Brown[1]

Developmental Testbed Center

1. National Center for Atmospheric Research, Boulder, CO
2. Earth System Research Laboratory, Boulder, CO
3. Cooperative Institute for Research in the Atmosphere, Colorado State University, Ft. Collins, CO

5/14/2010

DTC Developmental Testbed Center



Daily Severe/QPF Centerpoint

CAPS Vortex2 domain

Full

General Approach for

Based on DTC Model Evaluation Tools (MET)

2010 HWT Model Evaluations

- Models: 00Z and 12Z initializations (21Z and 09Z for SREF)
  - **CAPS Storm Scale Ensemble Forecast - 4km** - (all 26 **members**)
  - CAPS SSEF Ensemble Products - 4km - (15 members)
  - **HRRR – 3km**
  - NAM-218 – 12 km
  - Short Range Ensemble Forecast (SREF) Ensemble Products – 32 km
  - *Other models (NSSL, MMM, etc…) will be brought in for retrospective study*
- Variables:
  - **Reflectivity (REFC)**
  - Radar Echo Top Height of 18 dBZ contour ***(RETOP)***
  - 3 and 6 –hr Accum Precip Probability of Exceedance ***PROB(APCP_03>thresh) and PROB(APCP_06>thresh)***
  - *3 and 6-hr **Accum Precip (APCP_03) and (APCP_06)***
  - *Hourly probability of exceedance of reflectivity >40 dBZ : PROB(REFC>40)*

***Variables Evaluated in 2009**

FCST Field	Observation	Grid-Stat	MODE	Models
Prob of Exceed (0.5”, 1”, 2” over 3 and 6 hrs)	0.5”, 1”, 2” QPE over 3 and 6 hrs	**Brier Score, Decomp of Brier score, Area under ROC, Rel. Dia.**	None	Ensemble products from CAPS and SREF
50% Prob of Exceed (0.5”, 1”, 2” over 3 and 6 hrs)	0.5”, 1”, 2” QPE over 3 and 6 hrs	None	**MMI, Intersection Area, Area Ratio, Centroid Distance, Angle Difference, % Objects and Area Matched, 50th and 90th Percentile of Variable**	Ensemble products from CAPS and SREF
0.25”, 0.5”, 1.0”, 2” QPF over 3 and 6 hrs	0.25”, 0.5”, 1.0”, 2” QPE over 3 and 6 hrs	**GSS, CSI, FAR, PODY, FBIAS**	Same as above for 0.5” and 1.0”	CAPS members, CAPS ens mean, SREF ens mean, HRRR, NAM
Sim. CompositeRefl (20,30,40,50 dBZ)	Q2 Composite refl (20,30,40,50 dBZ)	**GSS, CSI, FAR, PODY, FBIAS**	Same as above for 30 dBZ initially 20,40 dBZ as resources allow	CAPS members, CAPS ens mean, HRRR, NAM
18 dBZ Echo Top (18, 25, 30, 35, 40, 45 kft)	Q2 18dBZ Echo Top (18, 25, 30, 35, 40, 45 kft)	**GSS, CSI, FAR, PODY, FBIAS**	Same as above for 25kFT initially 18 and 45 kFT as resources allow	CAPS members, CAPS ens mean, HRRR
Prob of 40dBZ echos	Q2 Composite reflectivity (40dBZ)	**Brier Score, Decomp of Brier score, Area under ROC, Reliability Diagram**	None	Ensemble products from CAPS and SREF
50% Prob of 40dBZ echos	Q2 Composite reflectivity (40dBZ)		See above	Ensemble products from CAPS

Verification Metrics

- Traditional Verification Metrics:
  - Categorical (Dichotomous) variables: GSS, CSI, FAR, PODY, FBIAS
- MODE Summary Metrics:
  - Derived values: Median of Maximum Interest (MMI), Total Interest
  - Attributes: Intersection Area, Area Ratio, Centroid Distance, Angle Difference, % Objects and Area Matched, Median Difference in 50th and 90th Percentile (forecast – observation objects)
- Probablistic Metrics:
  - Brier Score, Decomp of Briar score (reliability, resolution, uncertainty)
  - Area under Receiver Operating Characteristic curve (ROC)
  - *Reliability Diagram and ROC (*later in Experiment)*

Traditional Verification Metrics

Statistics for dichotomous variables

Contingency Table

Forecast at Threshold	Observed		
	Yes	**No**	
Yes	**Hits (YY)**	**False alarms (YN)**	YY + YN
No	**Misses (NY)**	**Correct rejections (NN)**	NY + NN
	YY + NY	YN + NN	**Total =** YY+YN+NY+NN

Table 1. Contingency table illustrating the counts used in verification statistics for dichotmous (e.g. Yes/No) forecasts and observations.

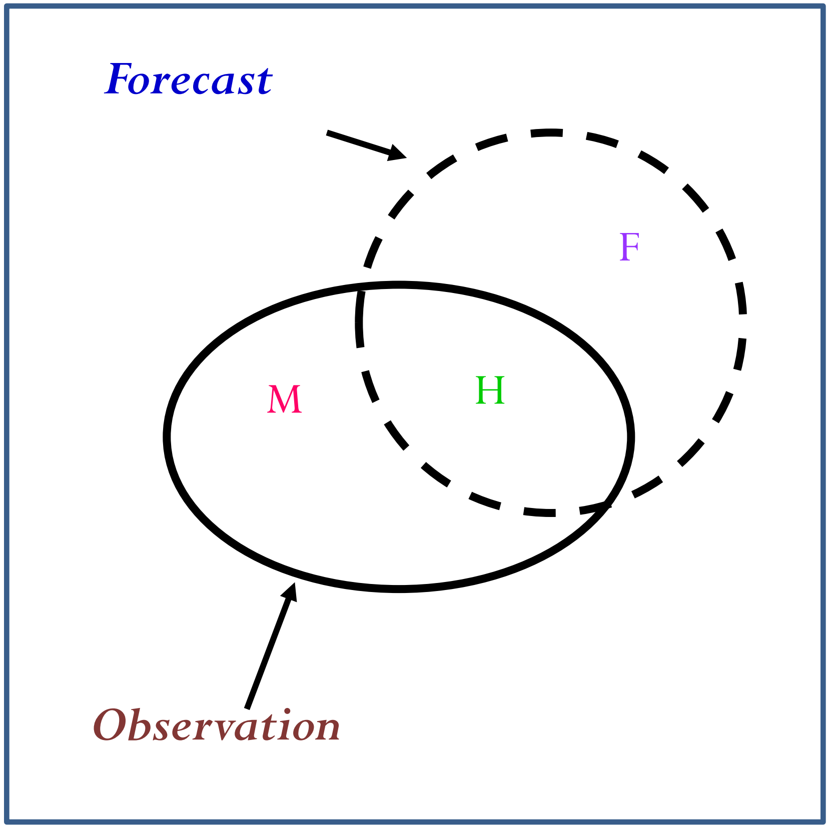
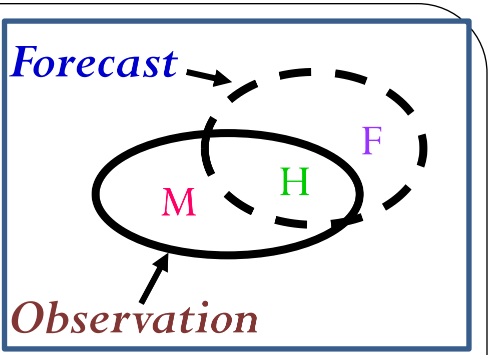


Figure 1. Diagram showing hits, misses, and false alarms for dichotomous forecast/observations.



Probability of Detection (PODY)

$$\frac{\#\text{Hits}}{\#\text{Hits} + \#\text{Misses}}$$

Range: 0 to 1. Perfect: 1

False Alarm Ratio (FAR)

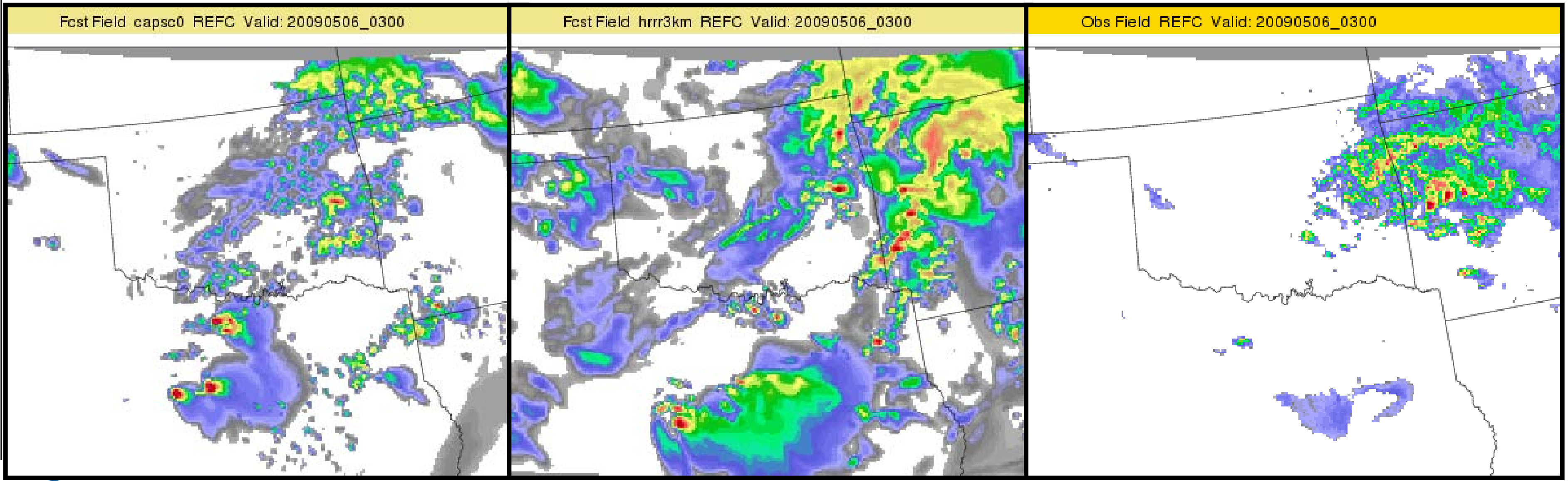
$$\frac{\#\text{False Alarms}}{\#\text{Hits} + \#\text{False Alarms}}$$

Range: 0 to 1. Perfect: 0

Base Rate (BASER)

$$\frac{\text{Observed Area}}{\text{Total Area}}$$

Range: 0 to 1. Complete Coverage: 1


Fcst Field capsc0 REFC Valid: 20090506_0300
Fcst Field hrrr3km REFC Valid: 20090506_0300
Obs Field REFC Valid: 20090506_0300

Lower FAR but also
Low PODY

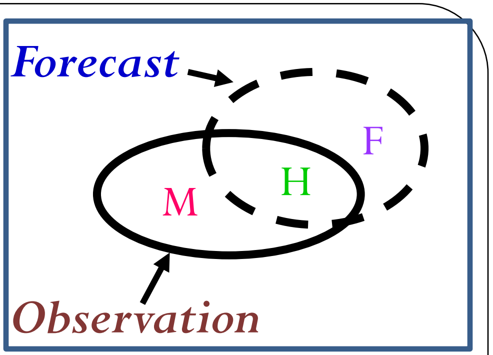
Higher FAR but also
Higher PODY

BASER for Low REFC ~ 0.33
BASER for High REFC ~0.01

Frequency Bias (FBIAS)

$$\frac{\textbf{Total Forecast Area}}{\textbf{Total Observation Area}}$$

Range: 0 to ∞. Perfect: 1



Critical Success Index (CSI)

$$\frac{\textbf{\#Hits}}{\textbf{\#Hits} + \textbf{\#Misses} + \textbf{\#False Alarm}}$$

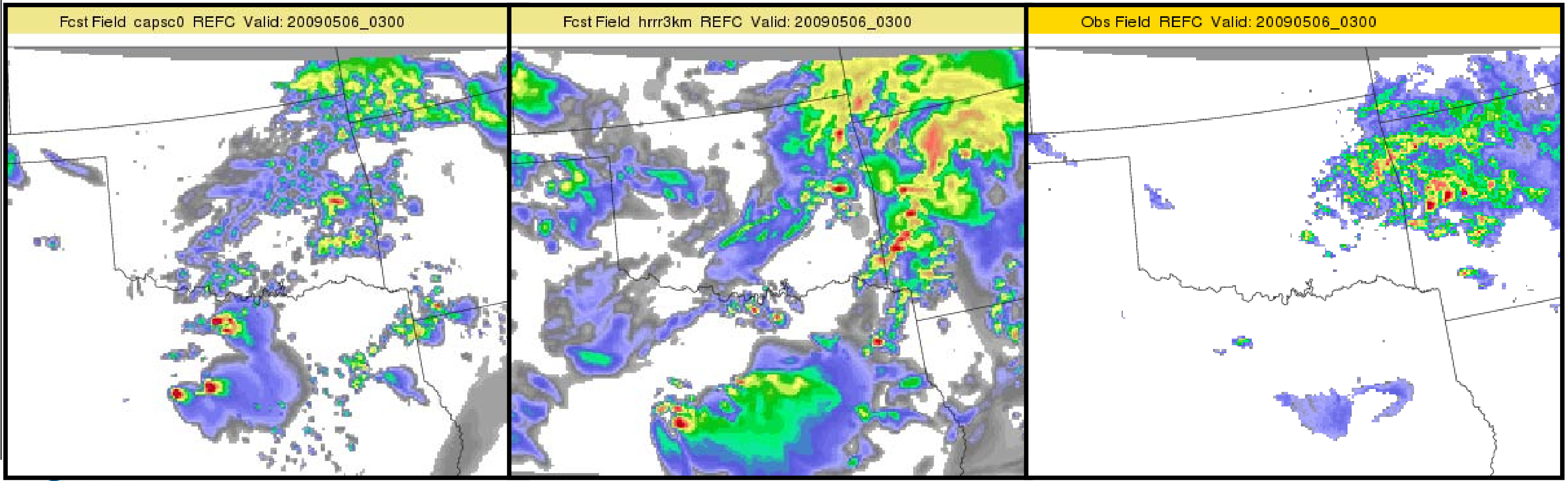
Range: 0 to 1. Perfect: 1

Gilbert Skill Score (GSS)

$$\frac{\textbf{\#Hits} - \textbf{\#Hits}_{\textbf{rand}}}{\textbf{\#Hits} + \textbf{\#Misses} + \textbf{\#False Alarm} - \textbf{\#Hits}_{\textbf{rand}}}$$

Range: -0.33 to 1. Perfect: 1

$$\textbf{\#Hits}_{\textbf{rand}} = \frac{(\textbf{Total Fcst Area})(\textbf{Total Obs Area})}{\textbf{Total Area}}$$



Fcst Field capsc0 REFC Valid: 20090506_0300

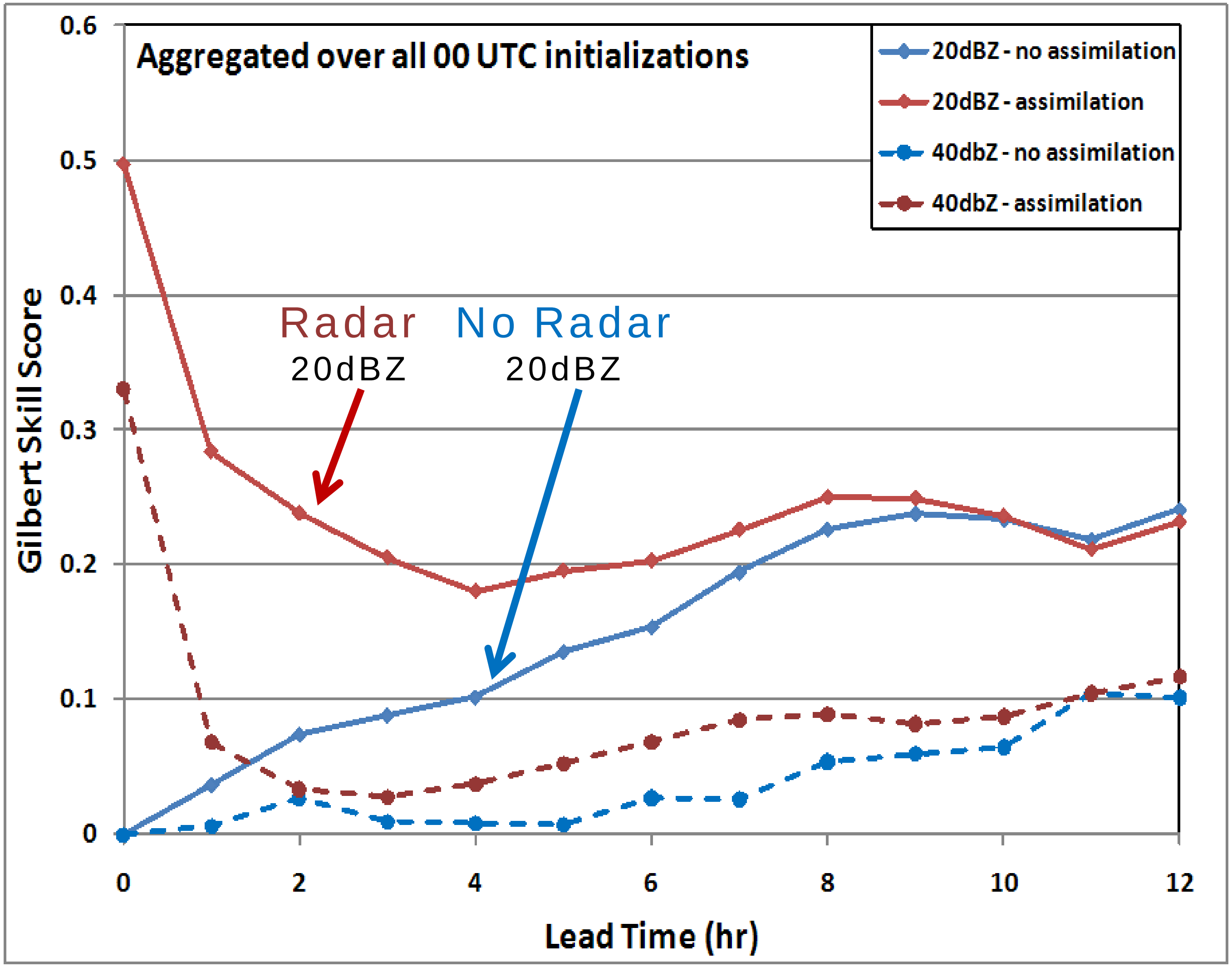
Fcst Field hrrr3km REFC Valid: 20090506_0300

Obs Field REFC Valid: 20090506_0300

Lower FBias but Higher GSS

Higher FBias – more Hits but prop. more False Alarms so lower GSS

Preliminary 2009 Results



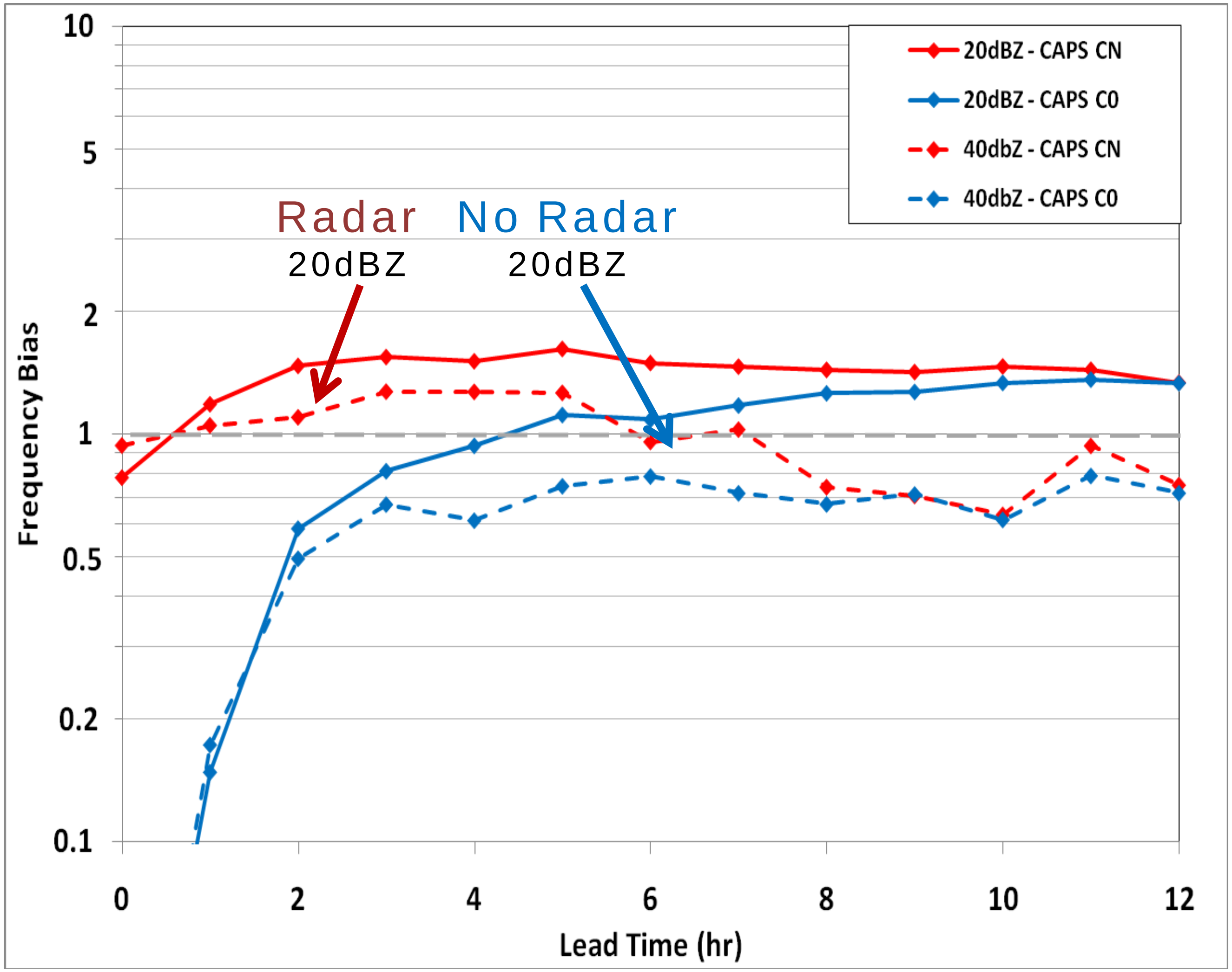
RESULS:

Radar assimilation appears to improve 0-6hr skill scores

Lack of clear difference in skill scores during 6-12 hr lead times suggests model physics taking over

Results were aggregated over Spring Experiment time period and the median values are plotted

Preliminary 2009 Results



Frequency Bias: Freq of fcst event / Freq of obs event

Assimilation
Over-fcst > 20 dBZ
Over-fcst > 40 dBZ
0-5 hr
Under-fcst>40 dBZ
6-12 hrs

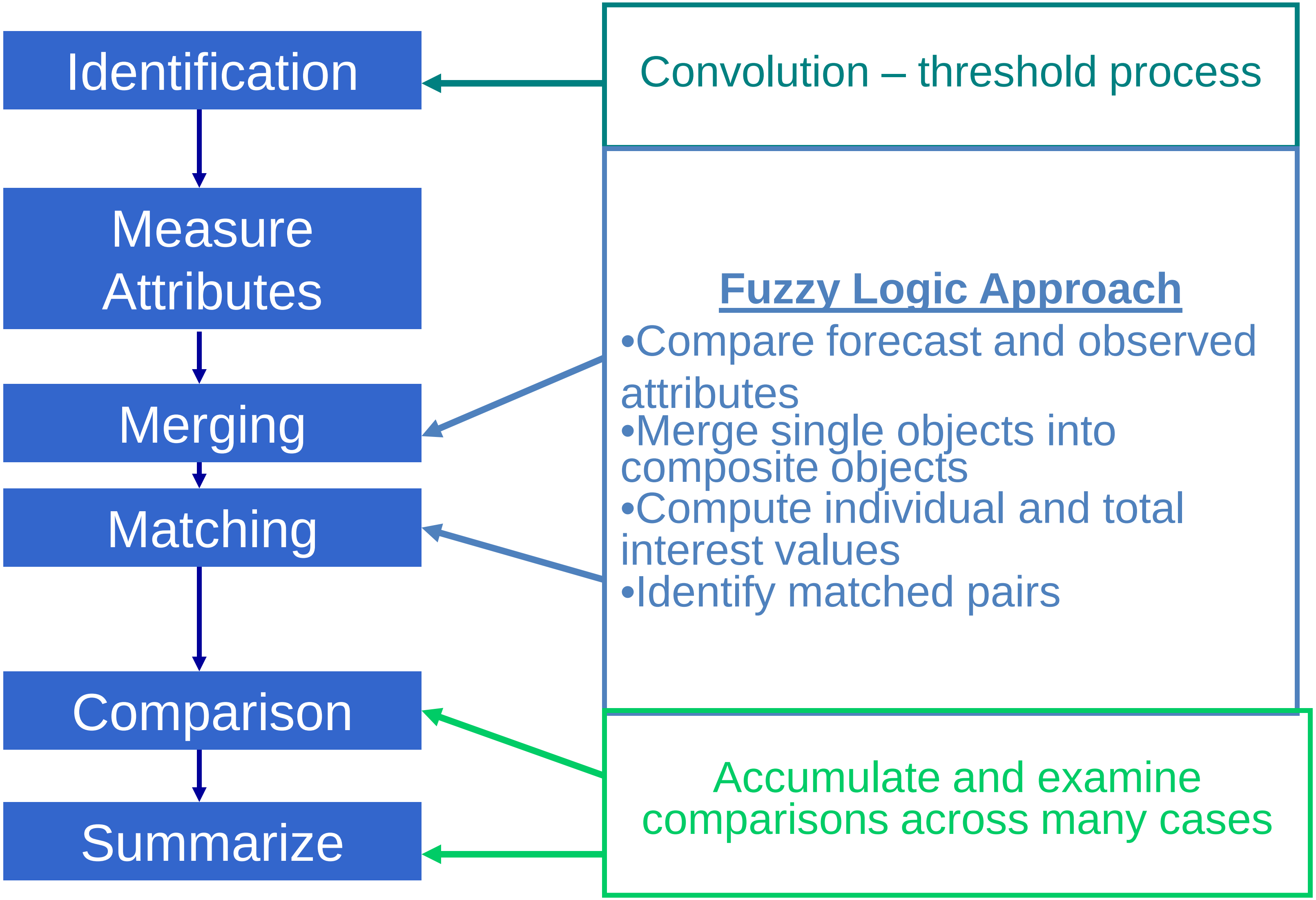
No assimilation
Under-fcst > 20 dBZ
0-4 hr
Over-fcst > 20 dBZ
0-5 hr
Under-fcst>40 dBZ

NOTE:
Lack of clear difference after lead time of 8hrs

Results were aggregated over Spring Experiment time period and the median values are plotted

Spatial Verification with MODE

MODE*: Object-based approach



Identification

Measure Attributes

Merging

Matching

Comparison

Summarize

Convolution – threshold process

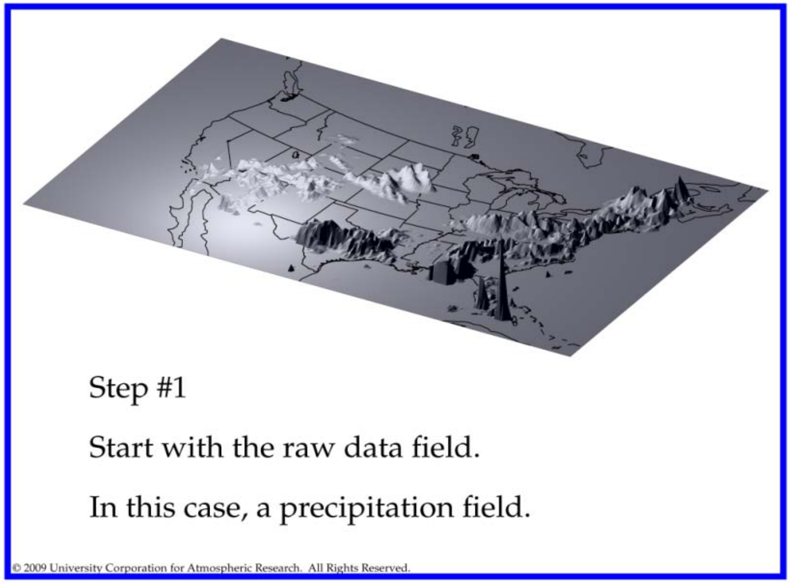
Fuzzy Logic Approach

- Compare forecast and observed attributes
- Merge single objects into composite objects
- Compute individual and total interest values
- Identify matched pairs

Accumulate and examine comparisons across many cases

*Method for Object-based Diagnostic Evaluation

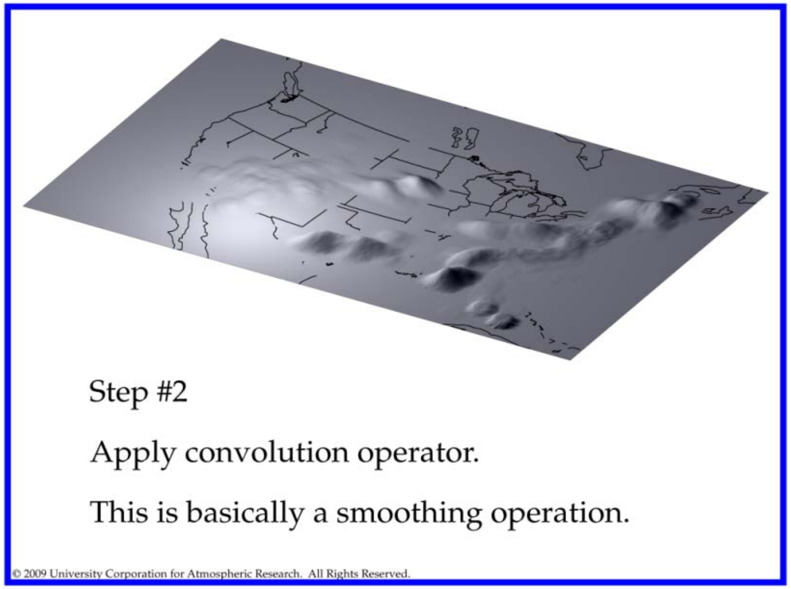
Object Definition



Step #1

Start with the raw data field.

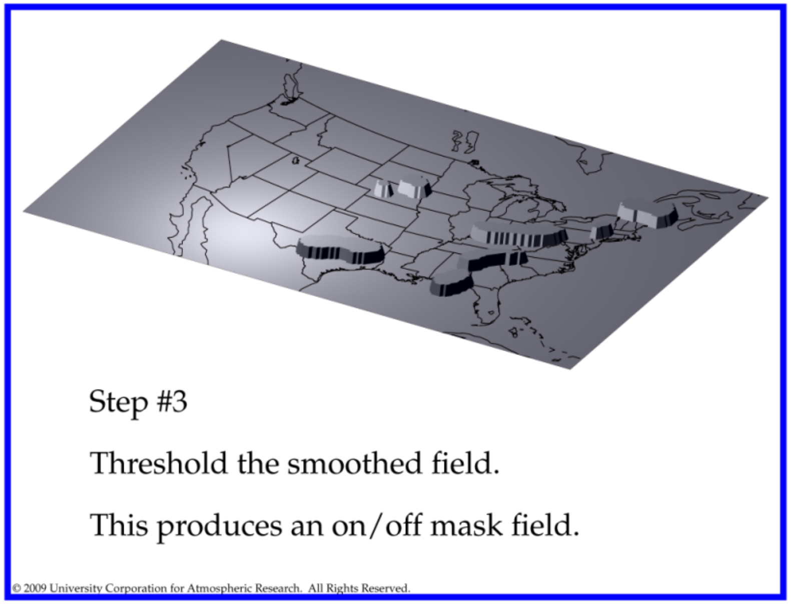
In this case, a precipitation field.



Step #2

Apply convolution operator.

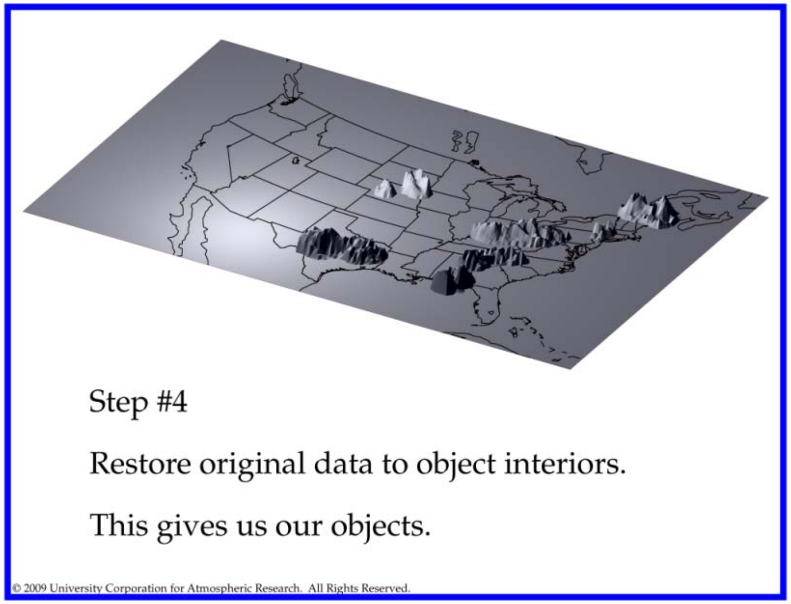
This is basically a smoothing operation.



Step #3

Threshold the smoothed field.

This produces an on/off mask field.

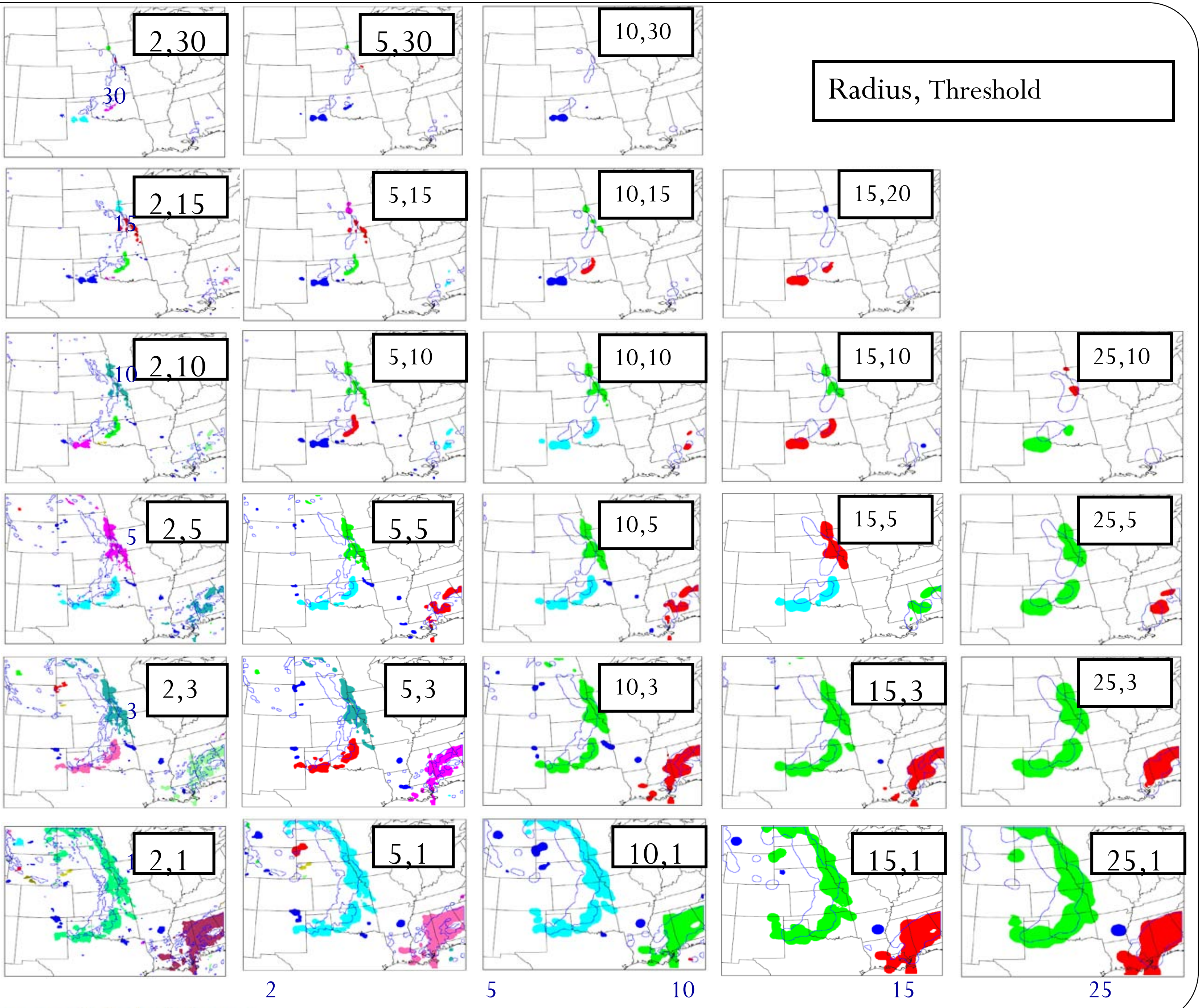


Step #4

Restore original data to object interiors.

This gives us our objects.

Threshold (in*100):



Radius (grid boxes):

MODE Attributes

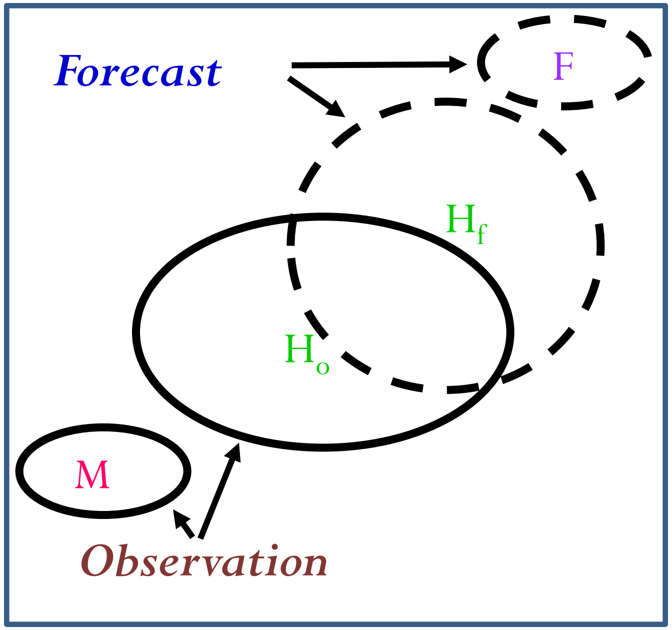
Intersection Area
Ratio of **intersection** area to **union** area

Area Ratio
Ratio of forecast to observation area

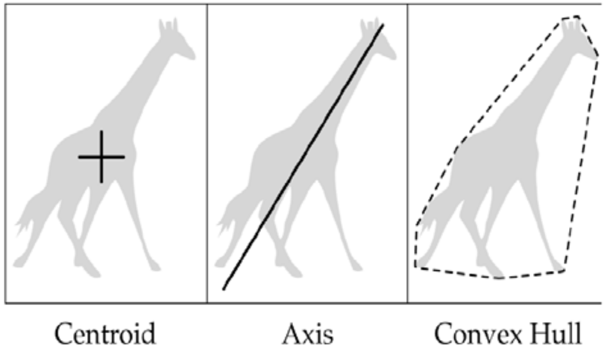
Centroid Distance
Distance between the **centroids**

Angle Difference
Difference between the axis angles of two objects

Percent Coverage
Percentage of evaluation area that is covered by observations and forecasts



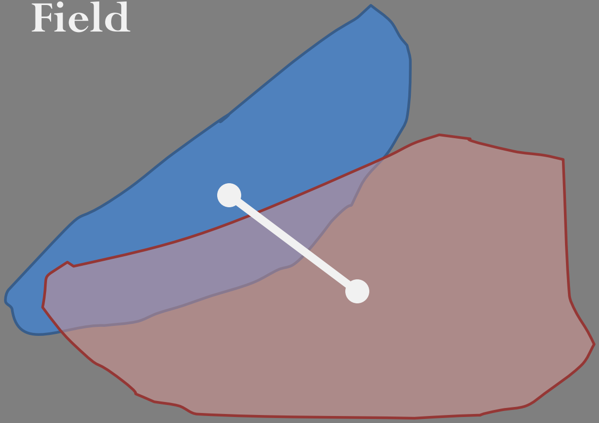
Example Single Attributes



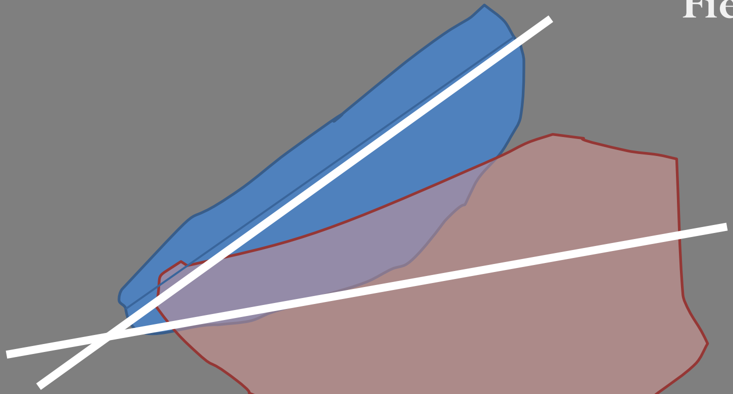
Use of Attributes of Objects defined by MODE

🔴 **Forecast Field**

🔵 **Observed Field**

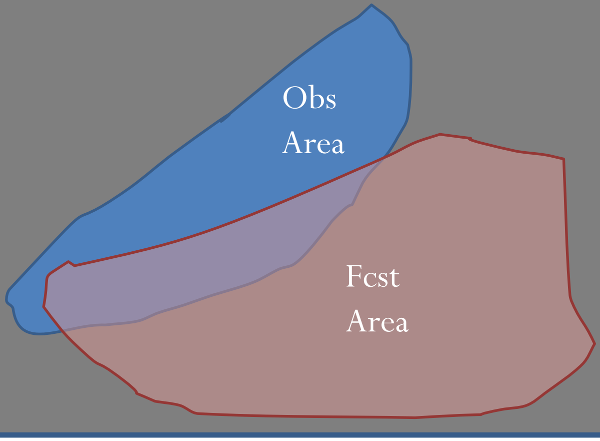


Centroid Distance: Provides a quantitative sense of spatial Displacement of AR core. ***Small is good***



Axis Angle: Provides an objective measure of how well the AR impact on terrain is captured. ***Small is good***

Area Ratio = $\frac{\text{Fcst Area}}{\text{Obs Area}}$

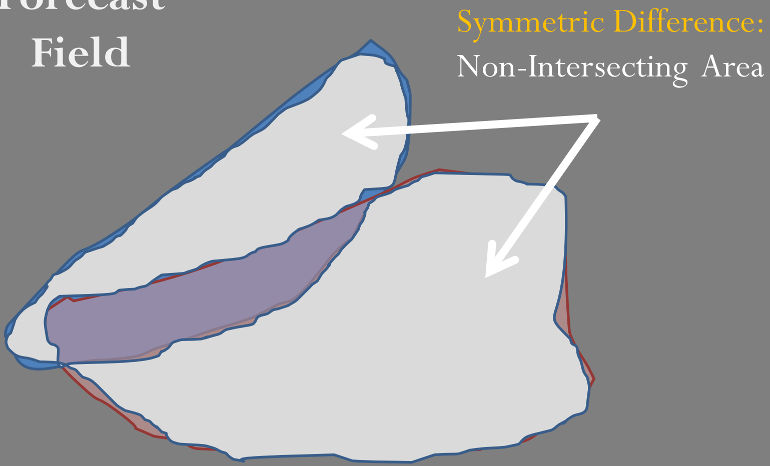


Area Ratio: Provides an objective measure of whether there is an over- or under-prediction of areal extent of AR. ***Close to 1 is good***

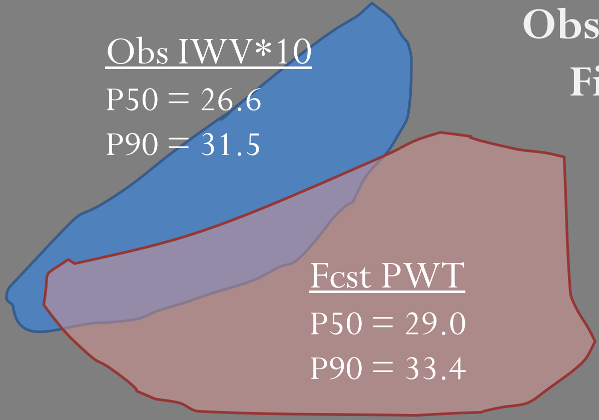
Use of Attributes of Objects defined by MODE

🟥 **Forecast Field**

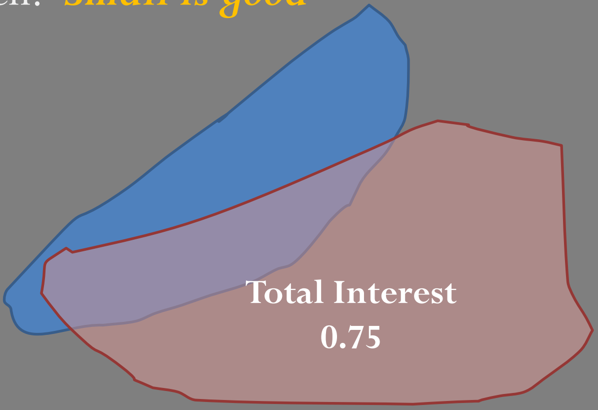
🟦 **Observed Field**



Symmetric Diff: May be a good summary statistic for how well Forecast and Observed objects match. ***Small is good***



P50/P90 Int: Provides objective measures of Median (50th percentile) and near-Peak (90th percentile) intensities found in objects. *Ratio close To 1 is good*

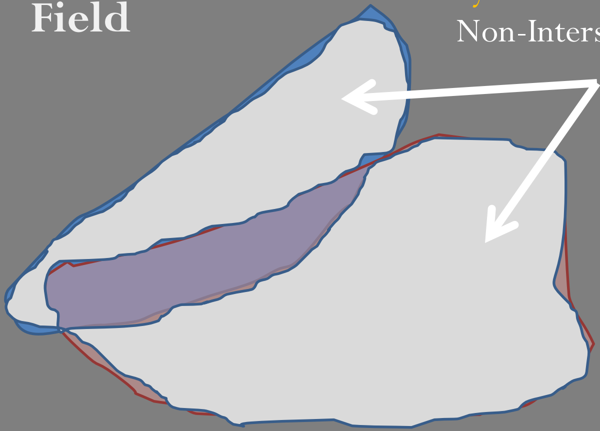


Total Interest: Summary statistic derived from fuzzy logic engine with user-defined Interest Maps for all these attributes plus some others. *Close to 1 is good*

Use of Attributes of Objects defined by MODE

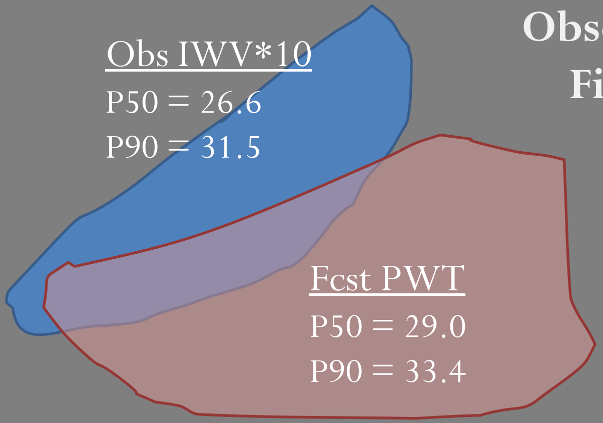
Forecast Field

Observed Field



Symmetric Difference:
Non-Intersecting Area

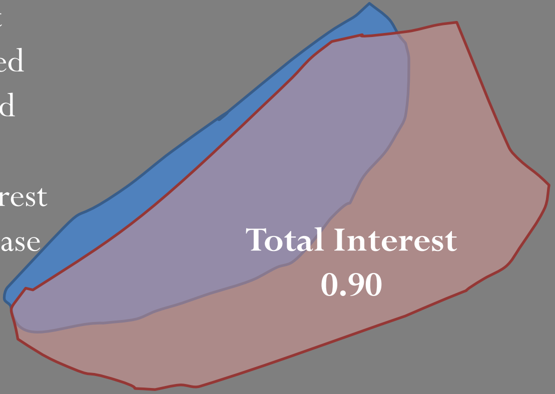
Symmetric Diff: May be a good summary statistic for how well Forecast and Observed objects match. ***Small is good***



P50/P90 Int: Provides objective measures of Median (50th percentile) and near-Peak (90th percentile) intensities found in objects.

Ratio close To 1 is good

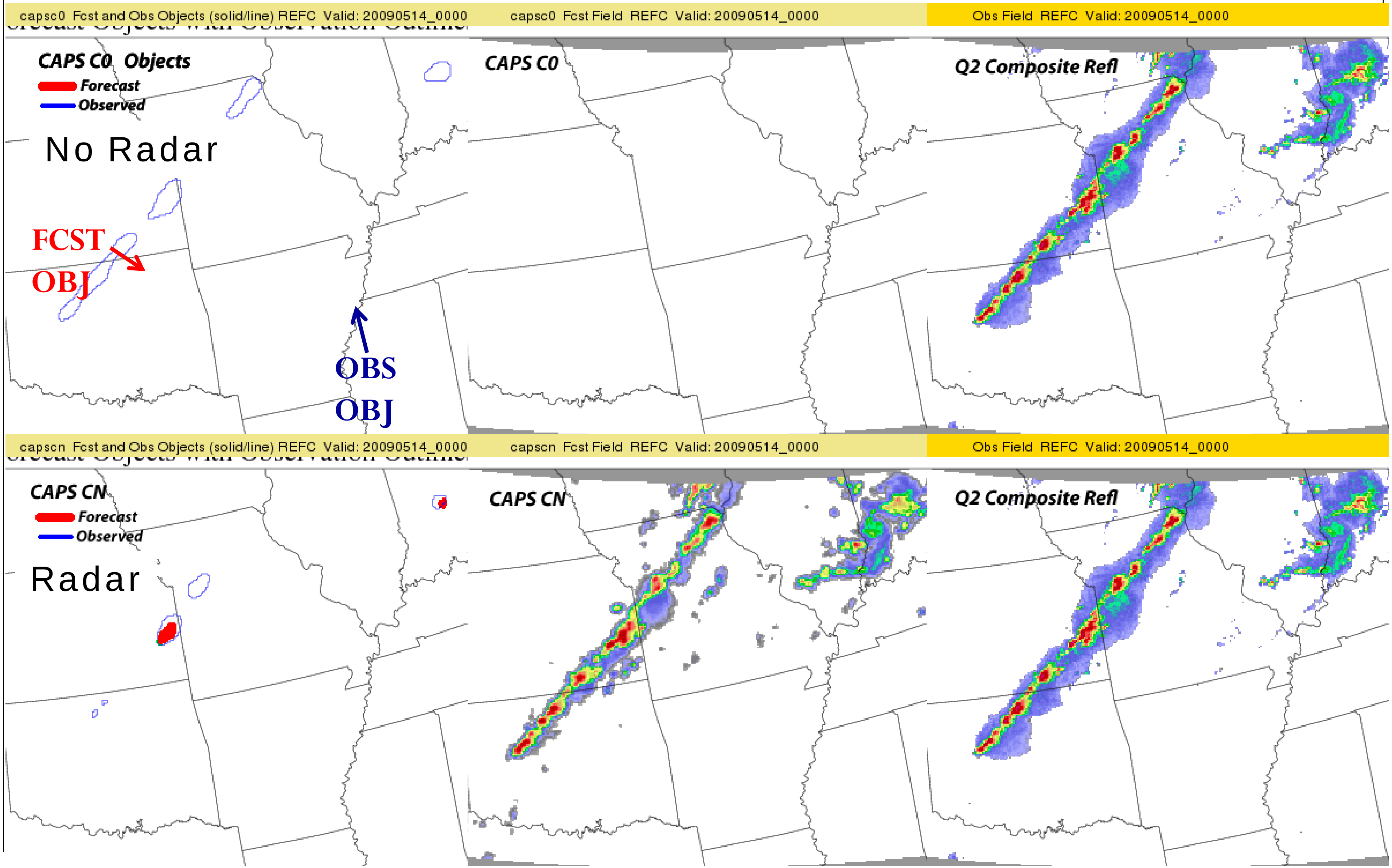
If forecast was rotated and moved North – Total Interest may increase



Total Interest: Summary statistic derived from fuzzy logic engine with user-defined Interest Maps for all these attributes plus some others.

Close to 1 is good

14 May 2009 Init: 00 UTC Spatial Thresh: 30dBZ


capsc0 Fcst and Obs Objects (solid/line) REFC Valid: 20090514_0000
capsc0 Fcst Field REFC Valid: 20090514_0000
Obs Field REFC Valid: 20090514_0000
CAPS C0 Objects
Forecast
Observed
No Radar
FCST OBJ
OBS OBJ
CAPS C0
Q2 Composite Refl
capscn Fcst and Obs Objects (solid/line) REFC Valid: 20090514_0000
capscn Fcst Field REFC Valid: 20090514_0000
Obs Field REFC Valid: 20090514_0000
CAPS CN
Forecast
Observed
Radar
CAPS CN
Q2 Composite Refl

Objects

Forecast Field

Observed Field

Calculate Total Interest and MMI

- Total Interest – uses Interest Map included in MODE config file
  - Allows user to weight importance of attributes
  - For example:
    - APCP – you could penalized for not hitting ACPC by ± 10% and not getting location within 10 grid points (40km)
    - RETOP – you could penalize for over predicting height by 10% but not under predicting height and not getting areal extent correct
    - REFC – you could heavily penalize for a underprediction of >20% and apply less penalty for < 20% error and not consider forecast that are more than 100km displaced
- **Once Total Interest is Calculated for each Object – a summary metric for entire grid is calculated - *Median of Maximum Interest***

Probabilistic Verification with Grid-Stat

Brier Score and Decomposition

- ***Brier score provides the user with a measure of*** *the magnitude of the probability forecast errors.*

BS = R*eliability* – R*esolution* + U*ncertainty* (Murphy 1973)

(see OpsPlan or MET Documentation for equation)

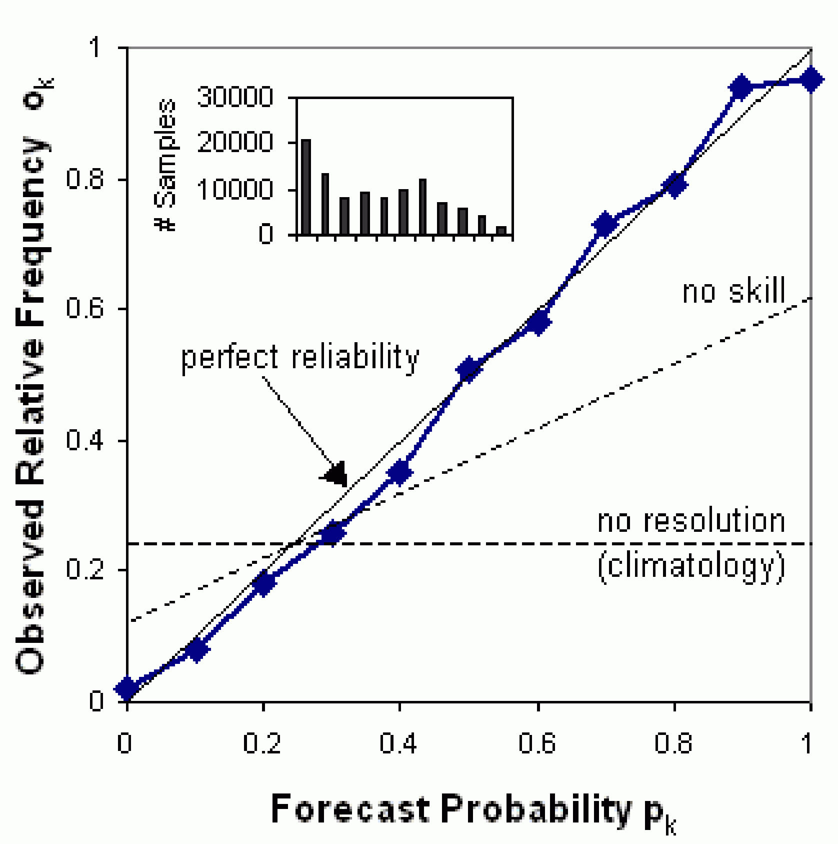
- It is suggested the user considers the homogeneity of the climatological mean when using the decomposition

- ***Answers the question:***

  What is the relative skill of the probabilistic forecast over that of climatology, in terms of predicting whether or not an event occurred?

- **Range:** 0 to 1, 1 indicates no skill w compared to the reference forecast.

  Perfect score: 0.



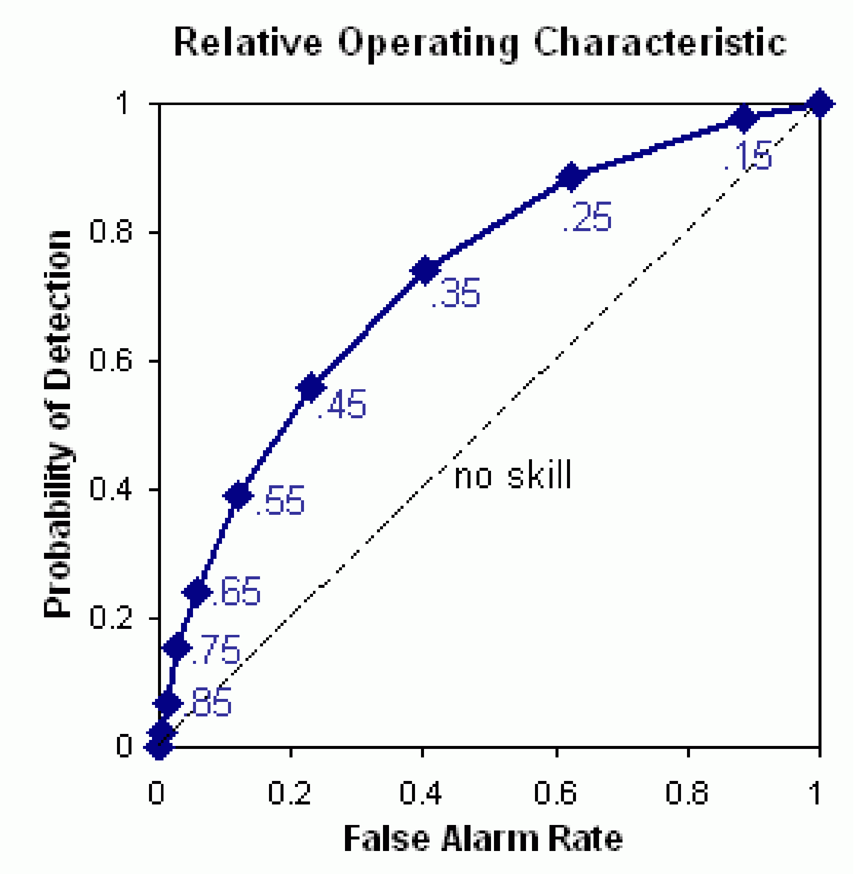
Area Under the ROC Curve

- *ROC*: **Perfect:** Curve travels from bottom left to top left of diagram, then across to top right of diagram. Diagonal line indicates no skill.

  Area under *ROC*: **Range:** 0 to 1, 0.5 indicates no skill.

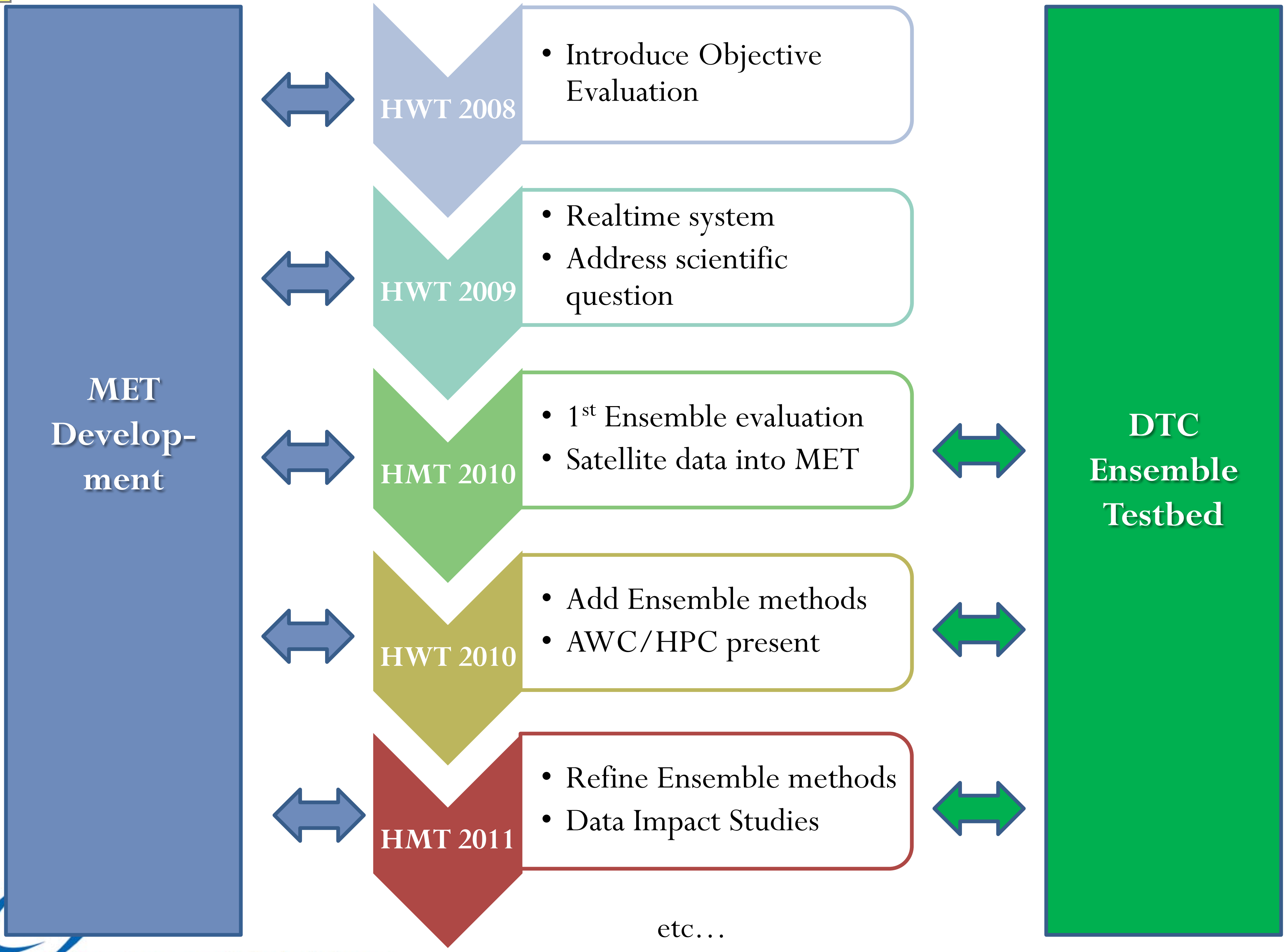
  Perfect score: 1

- ***Answers the question:*** *What is the ability of the forecast to discriminate between events and non-events?*



Just in case you were wondering…

YOUR ASSESSMENT OF DTC OBJECTIVE EVALUATION MATTERS…



MET Development

HWT 2008

- Introduce Objective Evaluation

HWT 2009

- Realtime system
- Address scientific question

HMT 2010

- 1st Ensemble evaluation
- Satellite data into MET

HWT 2010

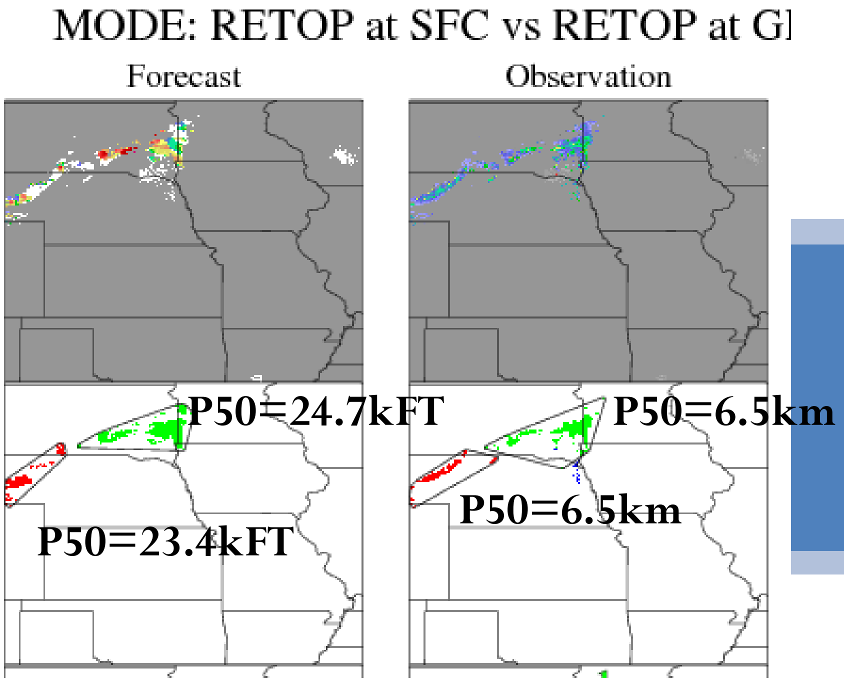
- Add Ensemble methods
- AWC/HPC present

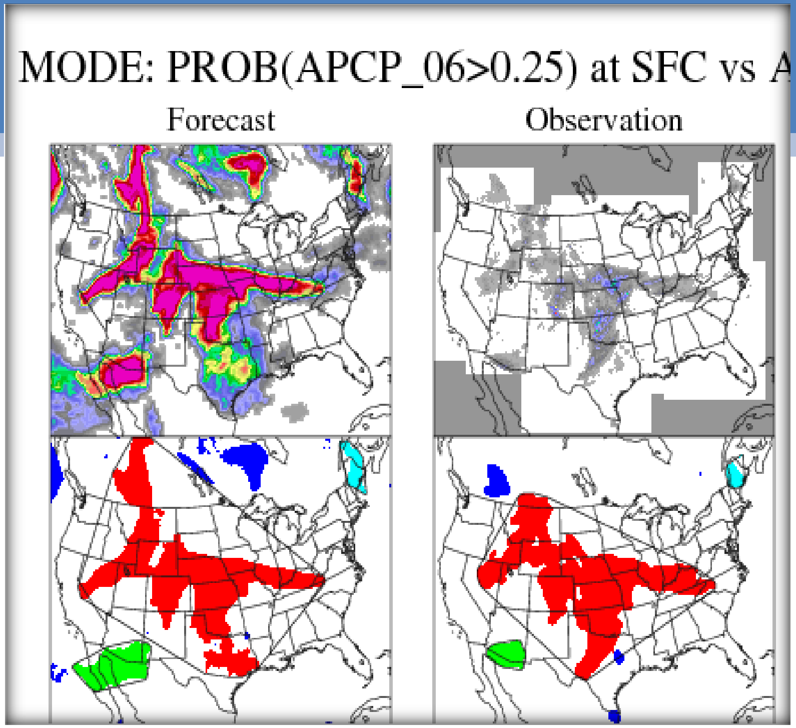
HMT 2011

- Refine Ensemble methods
- Data Impact Studies

etc…

DTC Ensemble Testbed





Thanks! Questions?

http://verif.rap.ucar.edu/eval/hwt/2010

Send E-mail to:

Tara Jensen - jensen@ucar.edu

DTC Developmental Testbed Center

***The Developmental Testbed Center is funded by the NOAA, AFWA and NCAR*

5/14/2010

Additional info on provided statistics and attributes…

Base Rate

$$\frac{\text{\#Hits} + \text{\#Misses}}{\text{Total Area}}$$

or

$$\frac{\text{Observed Area}}{\text{Total Area}}$$

Range: 0 to 1.

Depends on obs only.
Larger means more points for comparison and hence possibly more meaningful.

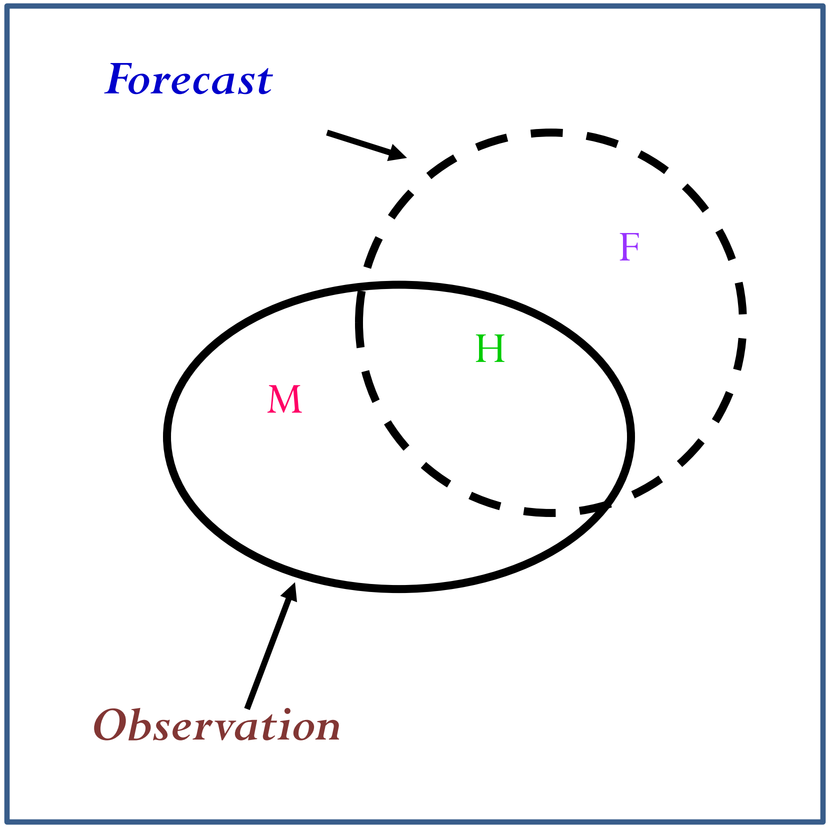


Figure 1. Diagram showing hits, misses, and false alarms for dichotomous forecast/observations.

False Alarm Ratio (FAR)

$$\frac{\#\text{False Alarms}}{\#\text{Hits} + \#\text{False Alarms}}$$

or

$$\frac{\text{Fcst Area where no Obs}}{\text{Total Forecast Area}}$$

Range: 0 to 1. Perfect: 0

Larger means less overlap area between fcst and obs. Should be used in conjunction with POD because ignores misses.

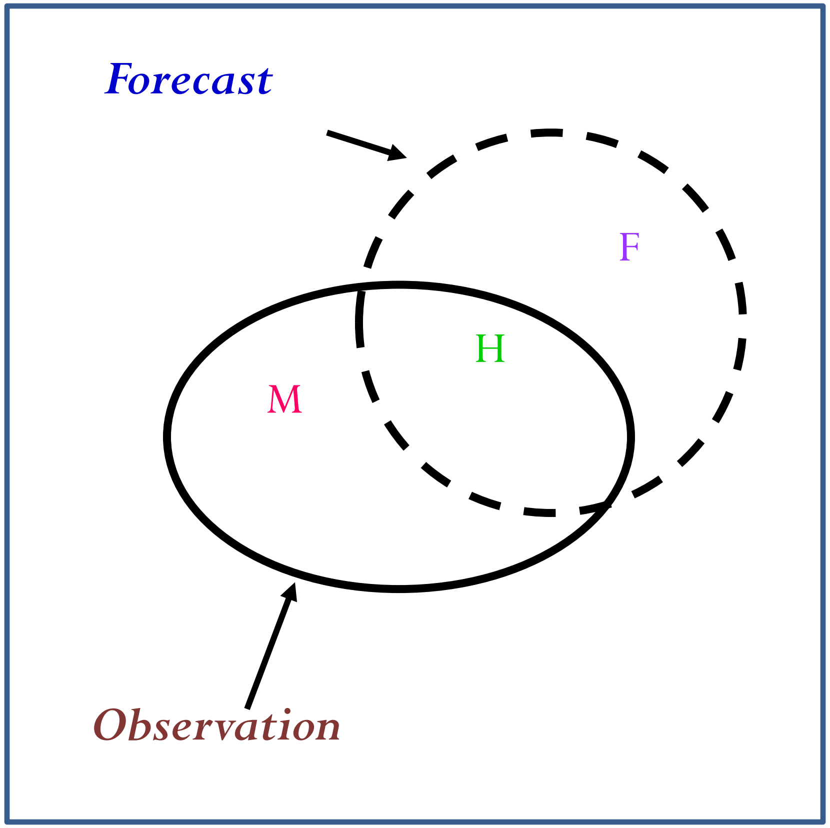
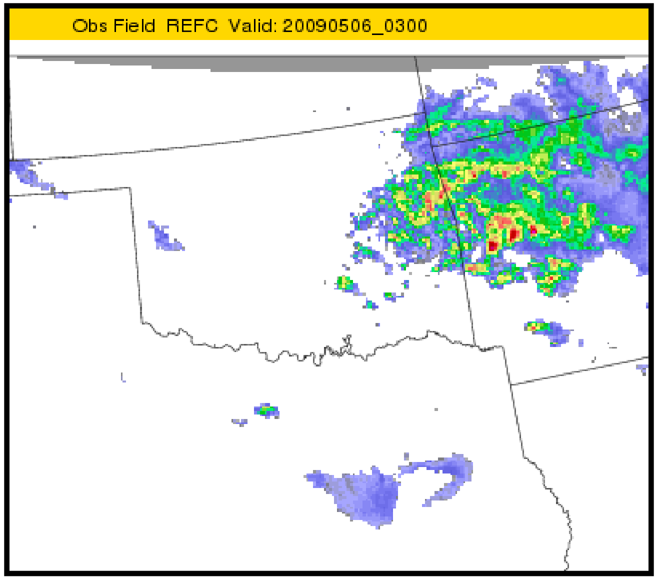
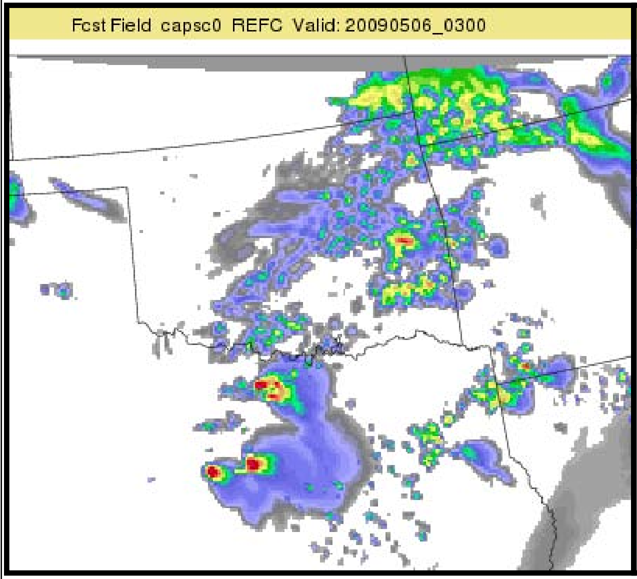
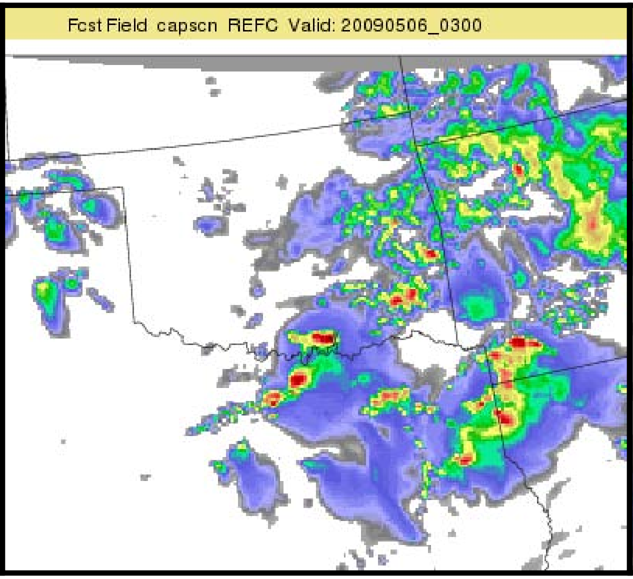


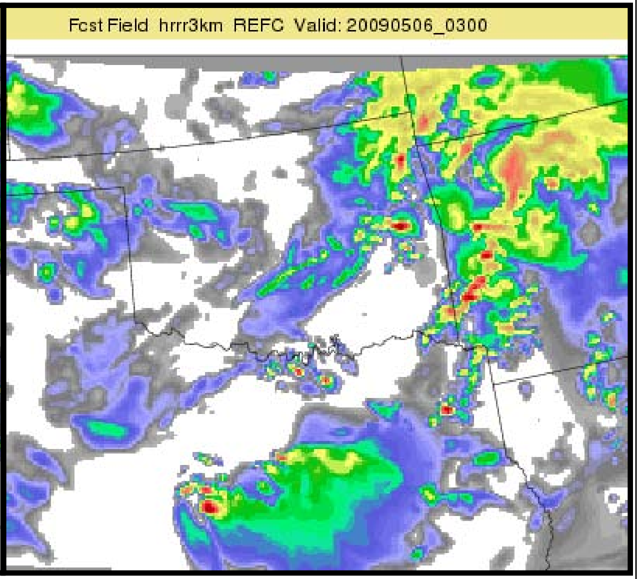
Figure 1. Diagram showing hits, misses, and false alarms for dichotomous forecast/observations.

Example

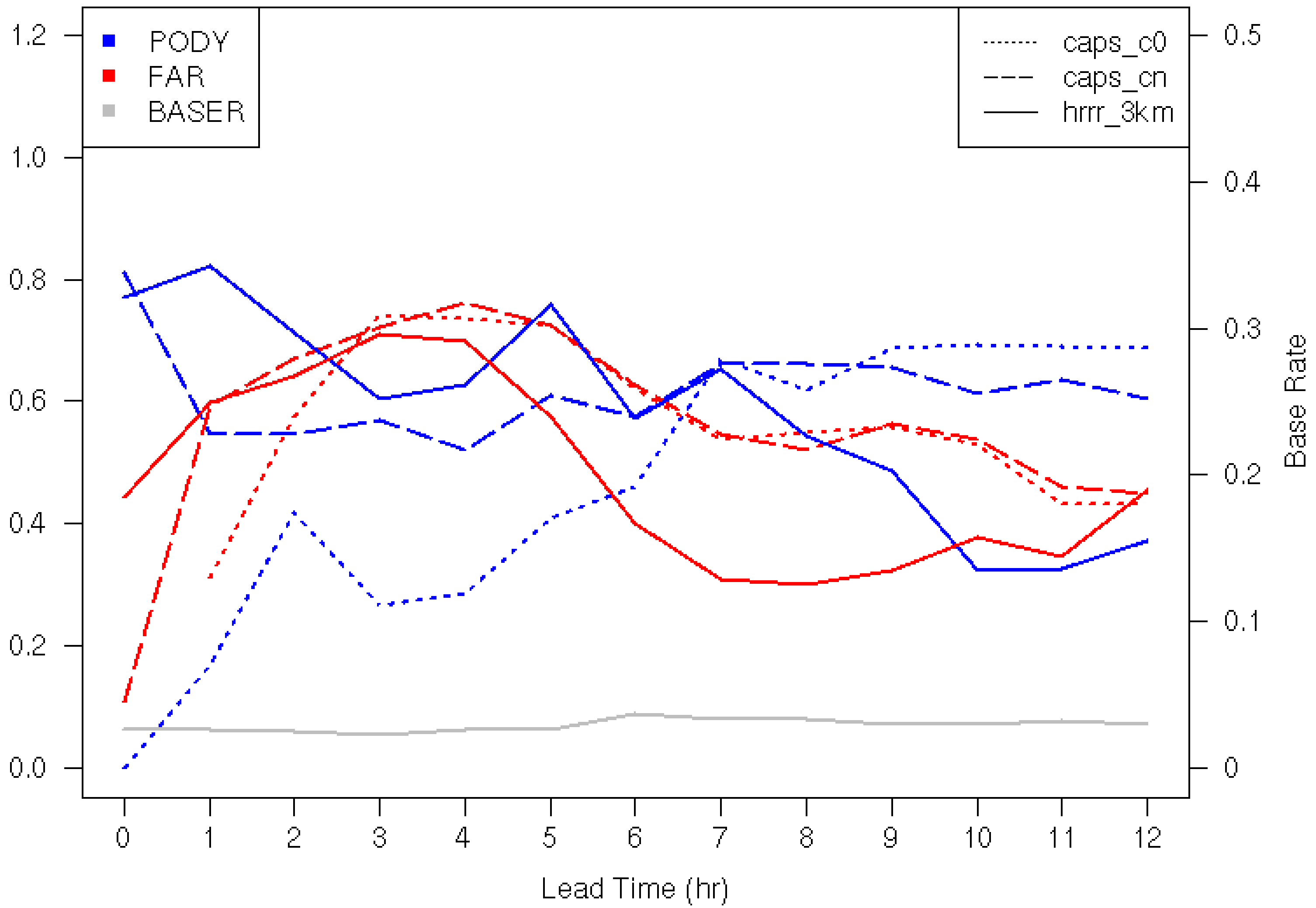

Obs Field REFC Valid: 20090506_0300


Fcst Field capsc0 REFC Valid: 20090506_0300

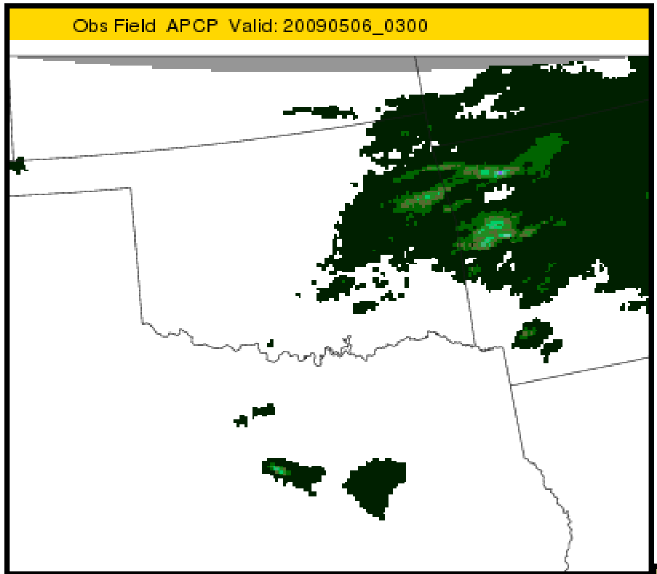

Fcst Field capscn REFC Valid: 20090506_0300

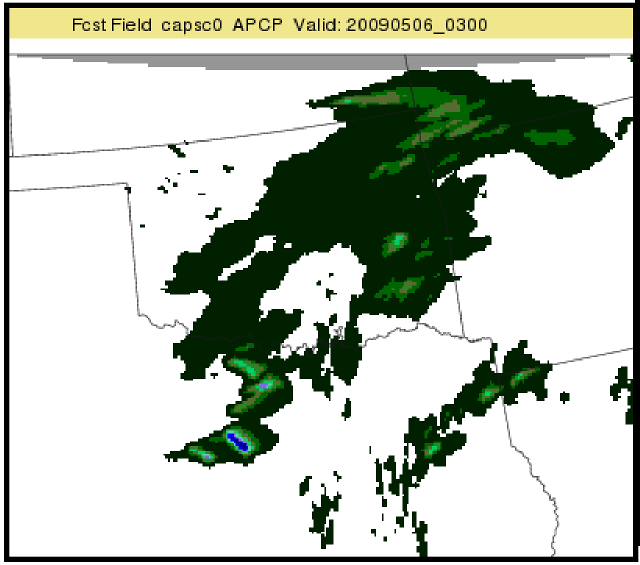

Fcst Field hrrr3km REFC Valid: 20090506_0300

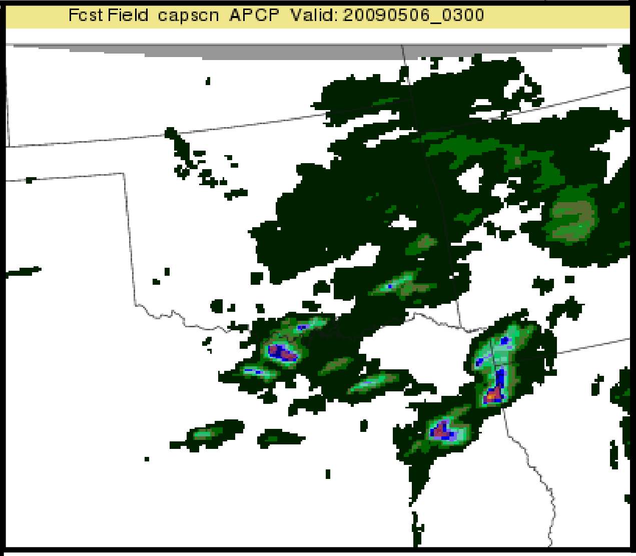
Initialization: 2009050600, Threshold: REFC>=20.000 dBZ

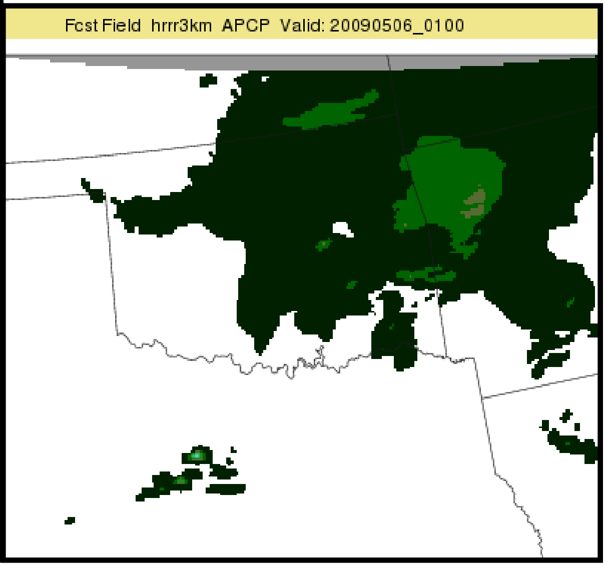


Example

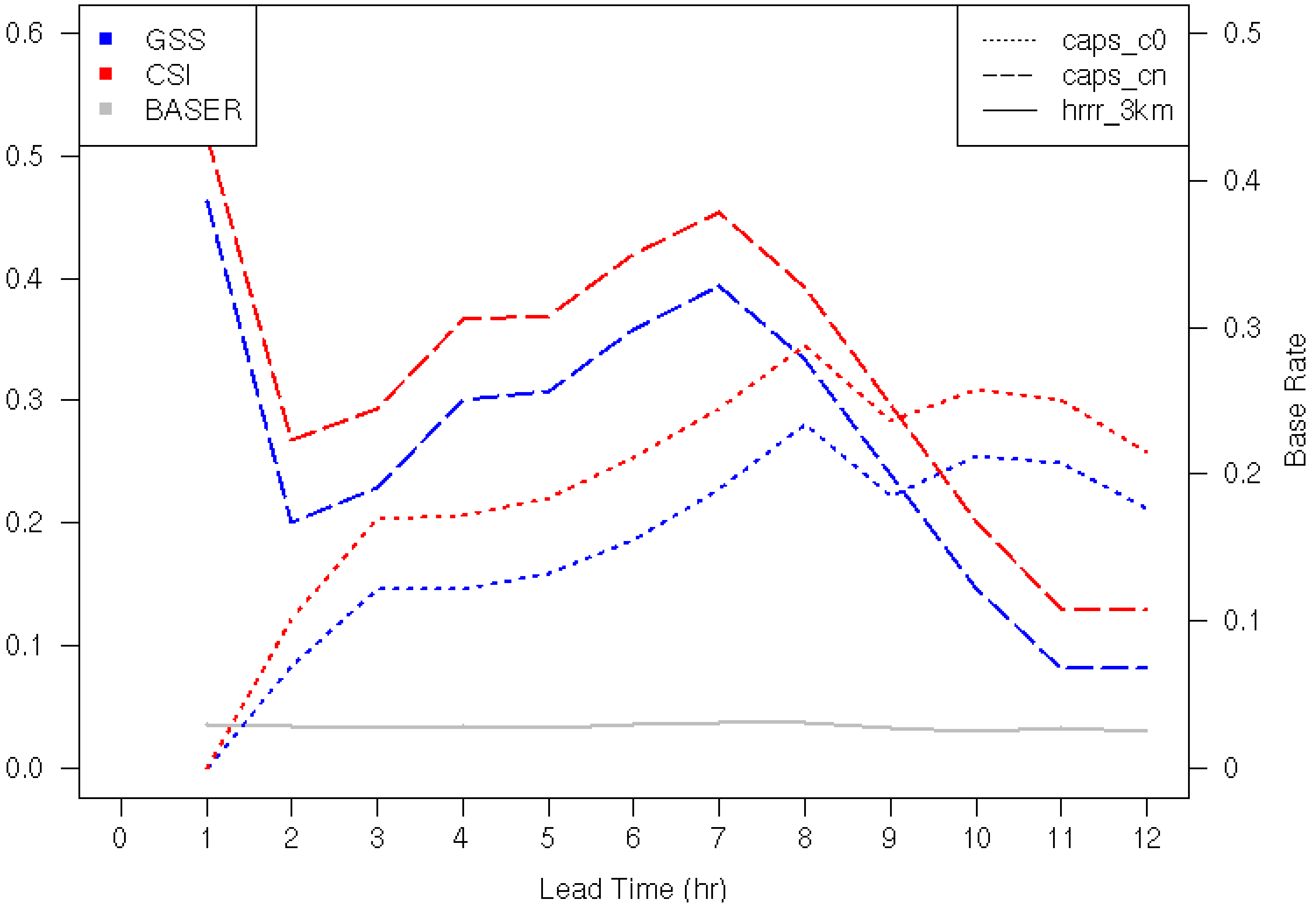

Obs Field APCP Valid: 20090506_0300


Fcst Field capsc0 APCP Valid: 20090506_0300


Fcst Field capscn APCP Valid: 20090506_0300


Fcst Field hrrr3km APCP Valid: 20090506_0100

Initialization: 2009050600, Threshold: APCP_01>=0.500 mm



Frequency Bias

$$\frac{\text{\#Hits} + \text{\#False Alarm}}{\text{\#Hits} + \text{\#Misses}}$$

or

$$\frac{\text{Total Forecast Area}}{\text{Total Observation Area}}$$

Range: 0 to ∞. Perfect: 1

<1: underforecast
>1: overforecast

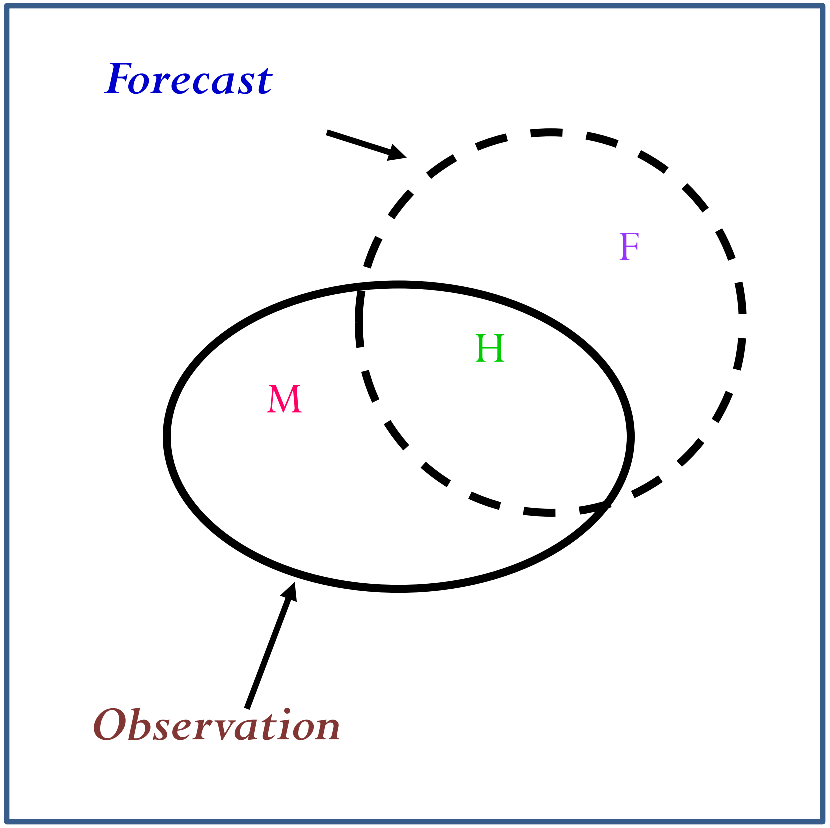


Figure 1. Diagram showing hits, misses, and false alarms for dichotomous forecast/observations.

Critical Success Index (CSI)

or Threat Score (TS)

$$\frac{\#\text{Hits}}{\#\text{Hits} + \#\text{Misses} + \#\text{False Alarm}}$$

or

$$\frac{\text{Overlap Area b/w Fcst and Obs}}{\text{Observed} + \text{Forecast Area}}$$

Range: 0 to 1.

It's a non-linear combination of POD and FAR. We recommend you look at POD and FAR also. Sensitive to hits, penalizes for misses and false alarms. Thought of as the accuracy when correct negatives have been removed from consideration.

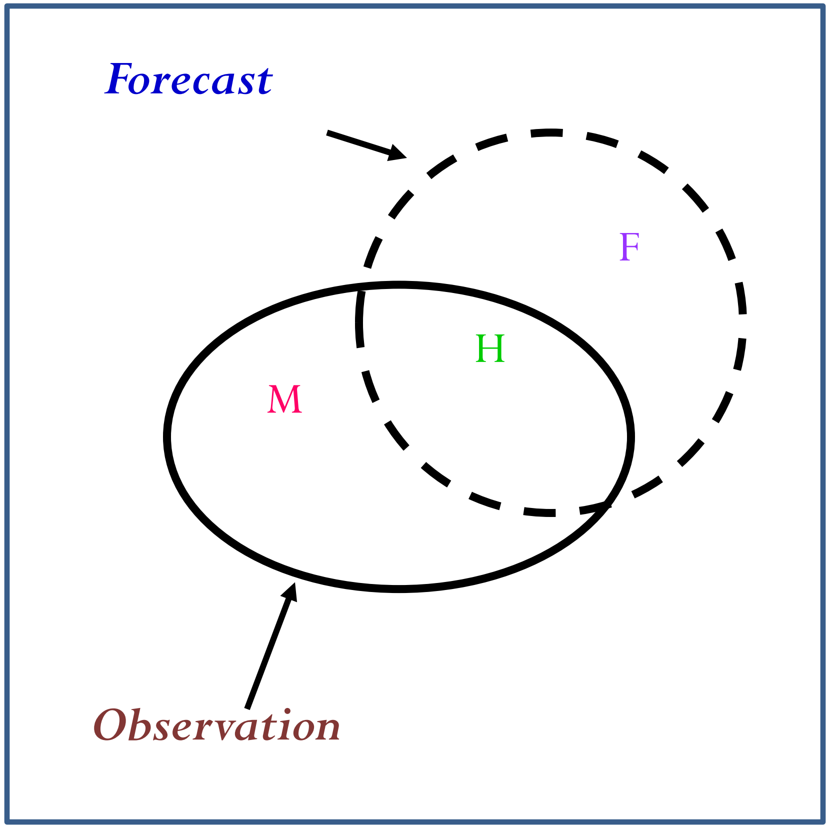


Figure 1. Diagram showing hits, misses, and false alarms for dichotomous forecast/observations.

Gilbert Skill Score (GSS)

Or Equitable Threat Score

$$\frac{\#\text{Hits} - \#\text{Hits}_{rand}}{\#\text{Hits} + \#\text{Misses} + \#\text{False Alarm} - \#\text{Hits}_{rand}}$$

where, $\#\text{Hits}_{rand} =$

$$\frac{(\text{Hits} + \text{False Alarm})(\text{Hits} + \text{Misses})}{\text{Total}}$$

or

$$\#\text{Hits}_{rand} = \frac{(\text{Total Fcst Area})(\text{Total Obs Area})}{\text{Total Area}}$$

Range: -0.33 to 1. Perfect: 1. No skill: 0.

Measures the fraction of observed and/or forecast events that were correctly predicted, adjusted for the frequency of hits that would be expected to occur simply by random chance.

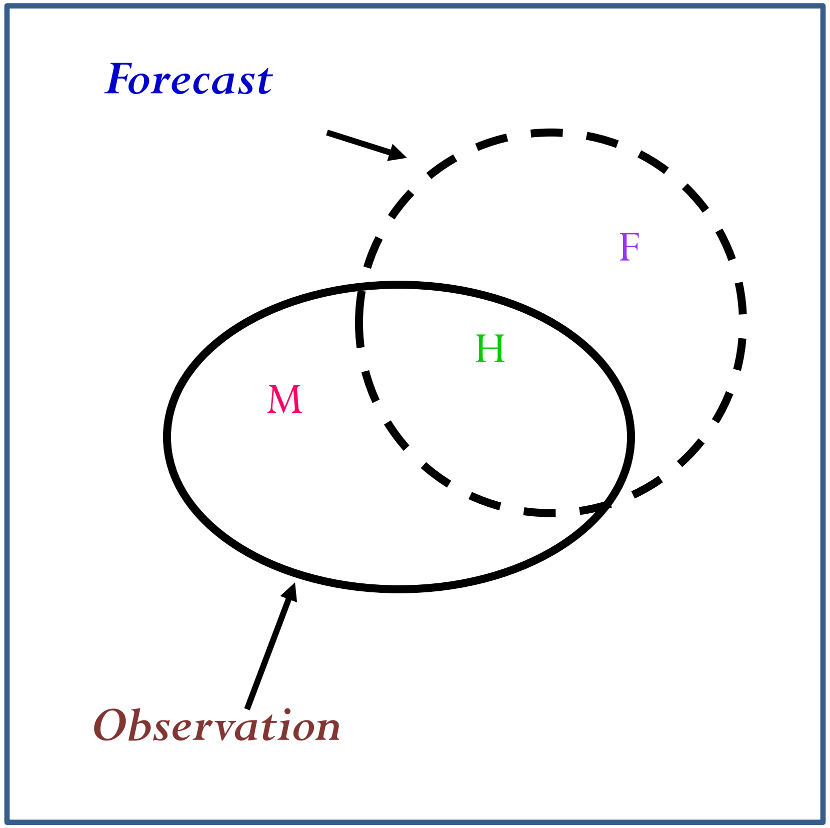
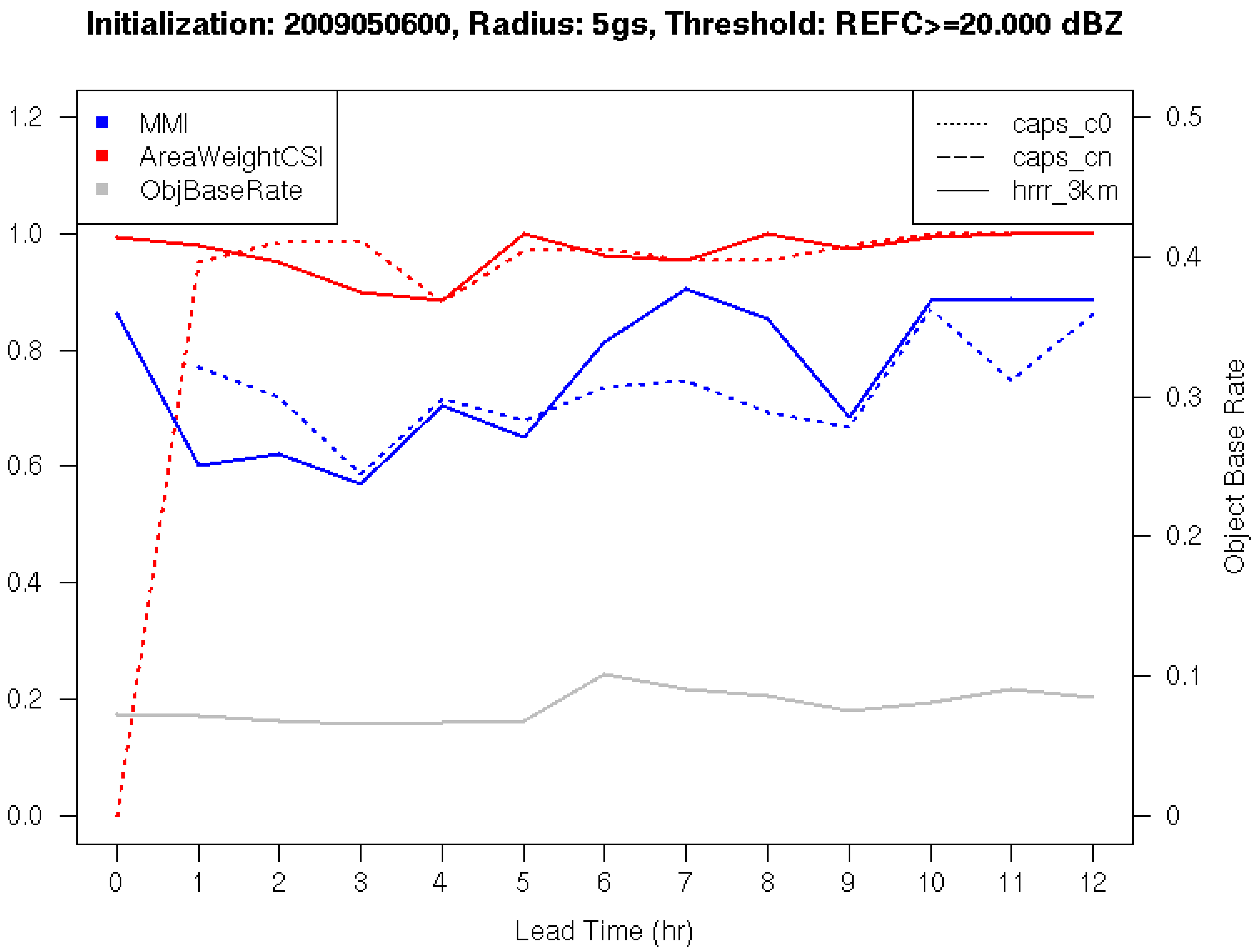
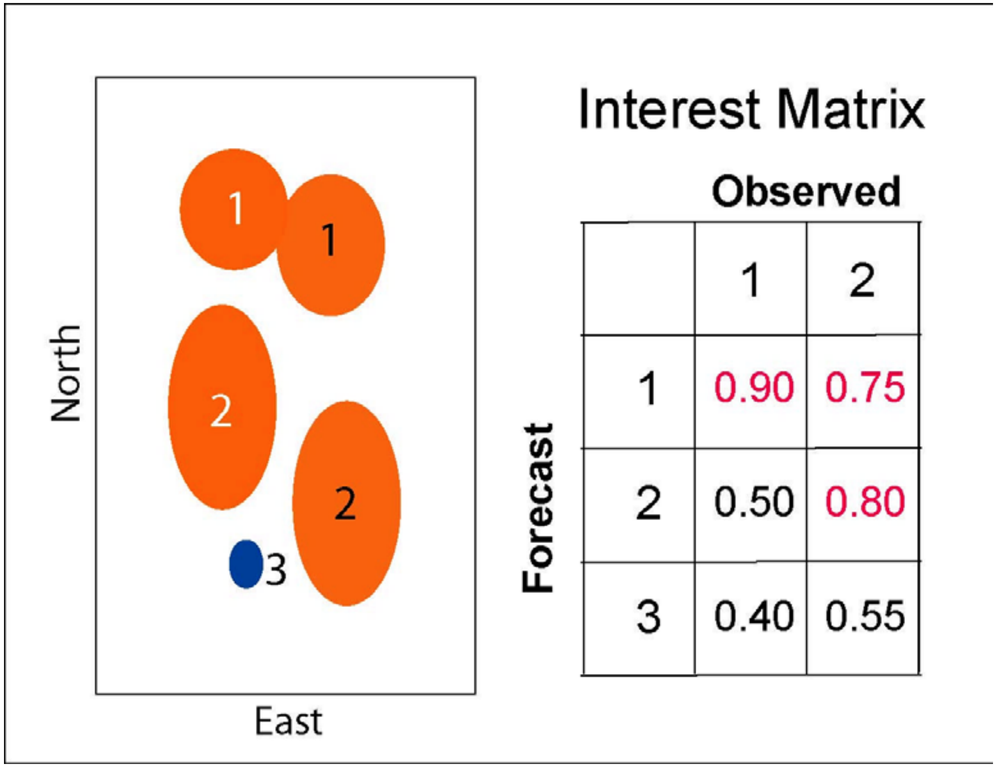


Figure 1. Diagram showing hits, misses, and false alarms for dichotomous forecast/observations.


Initialization: 2009050600, Radius: 5gs, Threshold: REFC>=20.000 dBZ
■ MMI
■ AreaWeightCSI
■ ObjBaseRate
caps_c0
caps_cn
hrrr_3km
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.5
0.4
0.3
0.2
0.1
0
Object Base Rate
0
1
2
3
4
5
6
7
8
9
10
11
12
Lead Time (hr)

MODE Summary Metrics

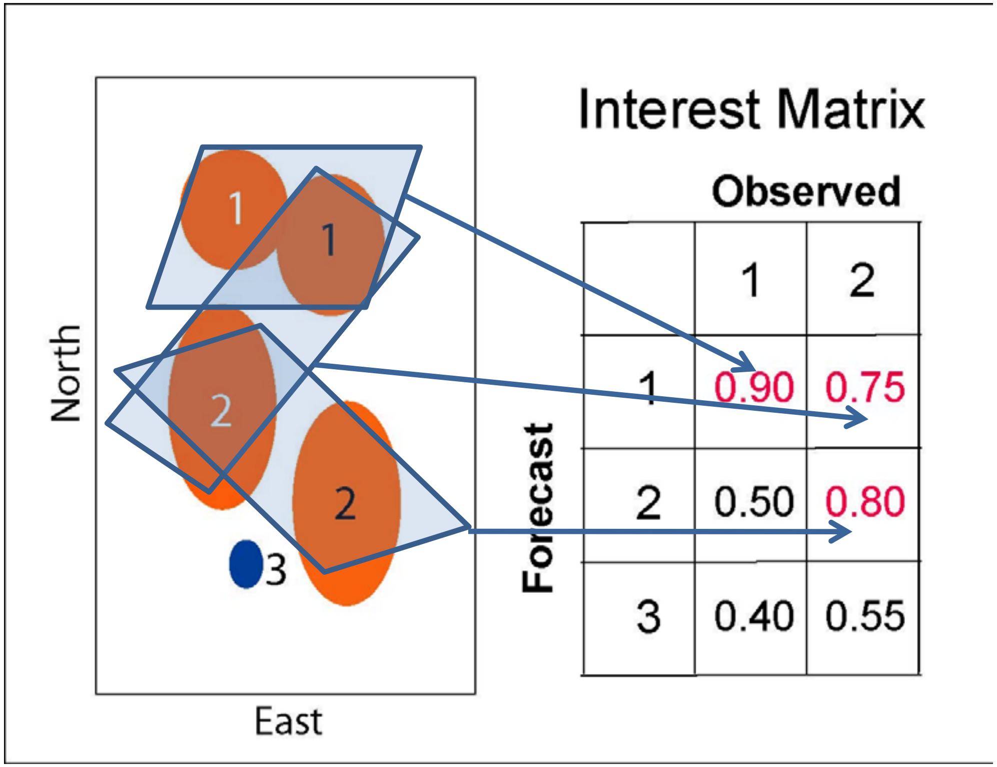
- Method for Object-based Diagnostic Evaluation (MODE)
  - User defined convolution radius (r) and precipitation/reflectivity threshold are used to identify objects
  - Objects are matched (associate objects in the fcst field with objects in the obs field) and merged (grouping of objects in the same field)
  - Forecast attributes that are used in the matching/merging process and to measure the quality of the forecast, include:
    - Object size
    - Distribution of intensity values
    - Orientation angle
    - Location



Interest Matrix

		Observed	
		1	2
Forecast	1	0.90	0.75
	2	0.50	0.80
	3	0.40	0.55

Figure 2. Schematic showing hypothetical forecast rain objects (black numerical labels) and observed rain objects (white numerical labels) with the corresponding interest matrix at right. Orange-shaded objects are matched whereas blue shading denotes no match. Total interest values greater than 0.7 are shown in red numbers in matrix. From Davis et al. (2009).

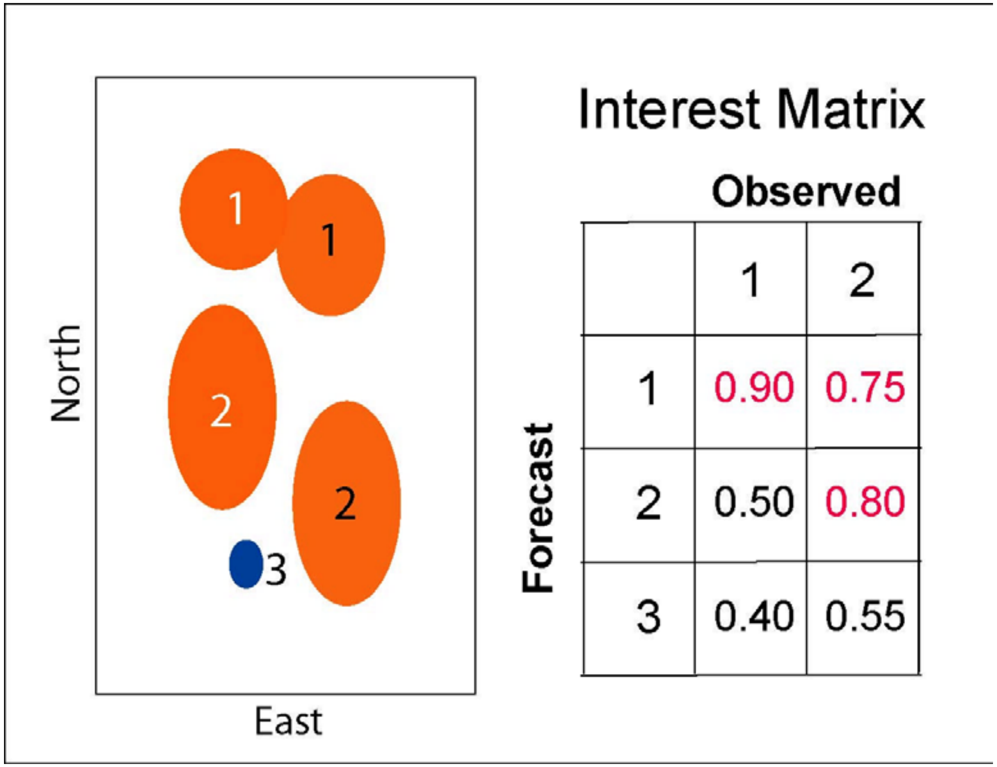


	1	2
1	0.90	0.75
2	0.50	0.80
3	0.40	0.55

Figure 2. Schematic showing hypothetical forecast rain objects (black numerical labels) and observed rain objects (white numerical labels) with the corresponding interest matrix at right. Orange-shaded objects are matched whereas blue shading denotes no match. Total interest values greater than 0.7 are shown in red numbers in matrix. From Davis et al. (2009).

To Summarize:

For forecast object 1, the maximum total interest is 0.90.
For forecast object 2, the maximum total interest is 0.80.
For forecast object 3, the maximum total interest is 0.55.
For observed object 1, the maximum total interest is 0.90.
For observed object 2, the maximum total interest is 0.80.



Interest Matrix

		Observed	
		1	2
Forecast	1	0.90	0.75
	2	0.50	0.80
	3	0.40	0.55

Figure 2. Schematic showing hypothetical forecast rain objects (black numerical labels) and observed rain objects (white numerical labels) with the corresponding interest matrix at right. Orange-shaded objects are matched whereas blue shading denotes no match. Total interest values greater than 0.7 are shown in red numbers in matrix. From Davis et al. (2009).

Median of Maximum Interest (MMI)

Considers the maximum total interest values associated with each forecast and observed object. From this set, the median value is computed.

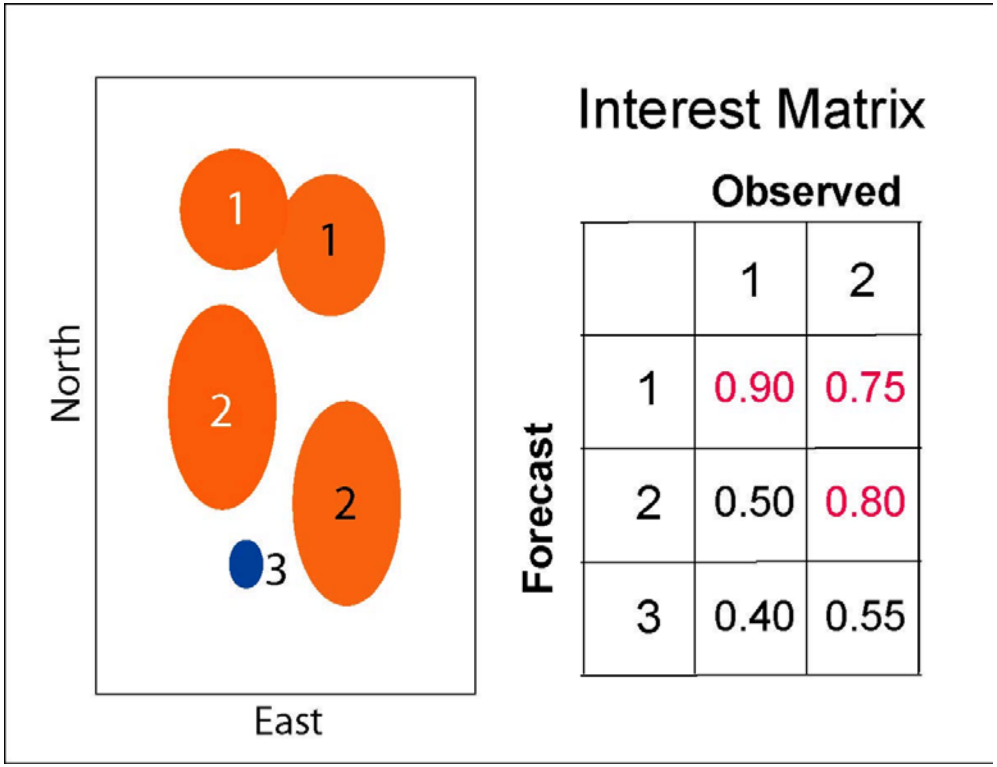
Range: 0 to 1.

<u>Example:</u>

For FO1, maximum Interest 0.90.
For FO2, maximum Interest is 0.80.
For FO3, maximum total interest is 0.55.
For OO1, maximum interest is 0.90.
For OO2, maximum interest is 0.80.

The median of those 5 numbers is 0.80, so MMI = 0.80.

Larger value suggests better match between all forecast and observed objects. Smaller value suggests objects do not match well or there are too many extra objects.



	1	2
1	0.90	0.75
2	0.50	0.80
3	0.40	0.55

Figure 2. Schematic showing hypothetical forecast rain objects (black numerical labels) and observed rain objects (white numerical labels) with the corresponding interest matrix at right. Orange-shaded objects are matched whereas blue shading denotes no match. Total interest values greater than 0.7 are shown in red numbers in matrix. From Davis et al. (2009).

Area-weighted CSI (AWCSI)

$$\frac{\#\text{Hits}}{\#\text{Hits} + \#\text{Misses} + \#\text{False Alarm}}$$

where

#Hits = Mean(H_o, H_f)

H_o = Matched Obs object area

H_f = Matched Fcst object area

#Misses = Unmatched Obs object area

#False Alarm = Unmatched Fcst object area

Range: 0 to 1. Perfect: 1. No skill: 0.

Hits based on object matching. Sensitive to hits, penalizes for misses and false alarms. Does not distinguish source of forecast error.

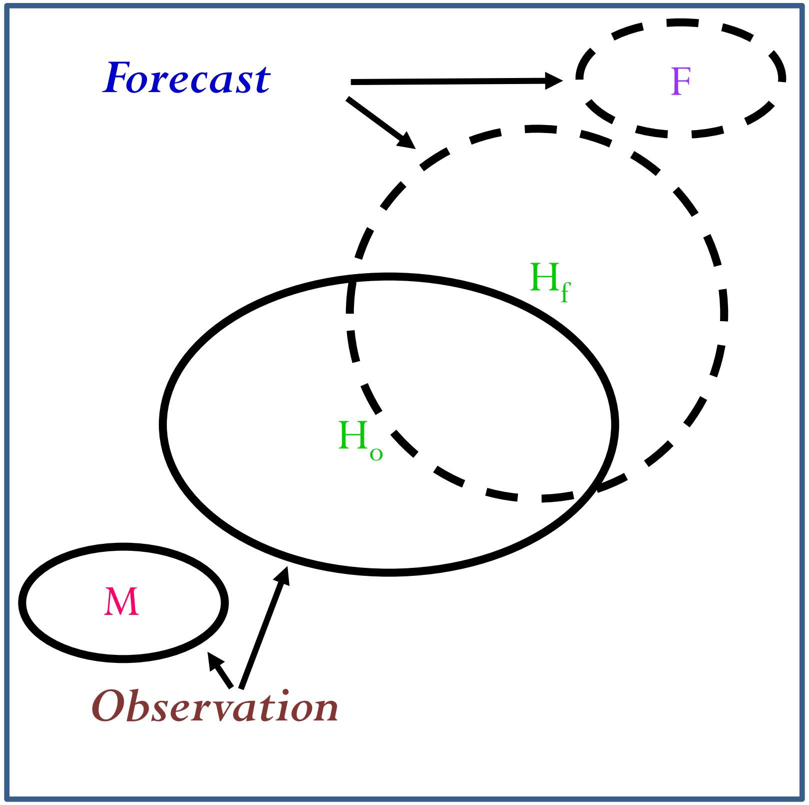


Figure 3. Diagram showing hits, misses, and false alarms for resolved forecast/observation objects.