

# Verification of Spatial Forecasts

Beth Ebert

Bureau of Meteorology Research Centre

Melbourne, Australia



# Issues in spatial verification

- What specific things do we want to get right?
- What verification scores and methods do we have at our disposal, and what do they really tell us about forecast quality?
- What are some of the advantages and disadvantages of these methods?



# Assumptions in spatial verification

1. The spatial forecast is, or can be put, on a regular grid
2. The verifying observations can be put onto the same grid as the forecast
3. The time difference between the forecast valid time and the observation time is acceptably small
4. The error in the “truth” data is much less than the difference between the forecast and the “truth”
5. Focus on one or more features of interest



# Does the spatial forecast look right?

- Is the feature of interest in the correct place?
- Does it have the correct size?
- Does it have the correct shape?
- Does it have the correct mean value?
- Does it have the correct maximum value?
- Does it have the correct spatial variability?



# Spatial verification methods

- Visual ("eyeball") verification
- Continuous statistics
- Categorical statistics
- Joint distributions

"standard"

---

"scientific" or "experimental"

- Scale decomposition methods
- Entity-based methods
- Event-oriented methods

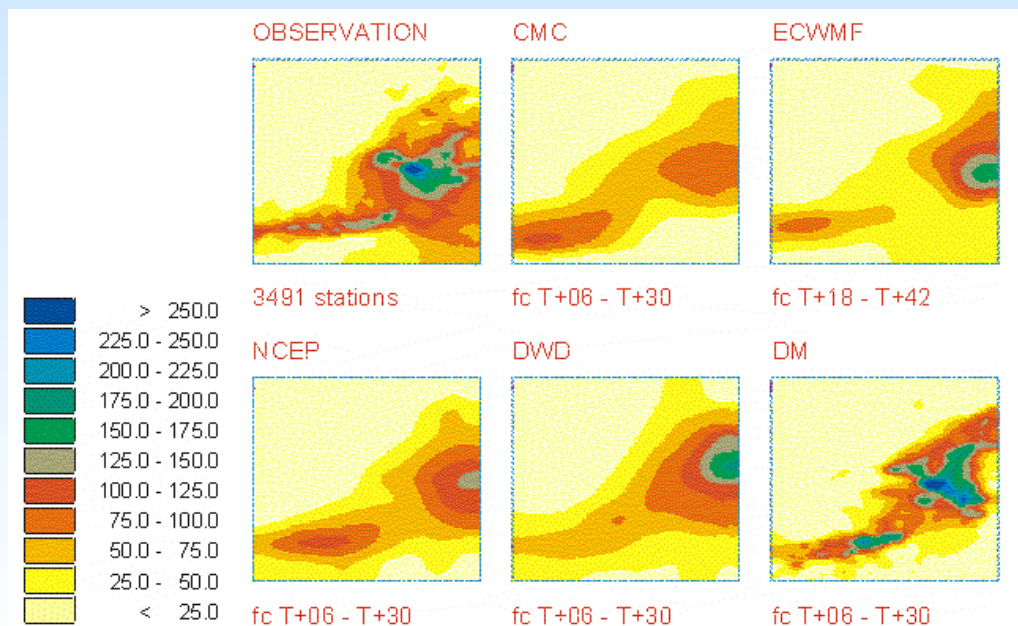


# Visual ("eyeball") verification

Visually compare maps of forecast and observations

**Advantage:** "A picture tells a thousand words..."

**Disadvantages:** Labor intensive, not quantitative, subjective



Verifies this attribute?	
Location	✓
Size	✓
Shape	✓
Mean value	✓
Maximum value	✓
Spatial variability	✓



# Continuous verification statistics

Measure the correspondence between the *values* of the forecasts and observations at the gridpoints

Examples:

- mean error (bias)
- mean absolute error
- root mean squared error
- linear error in probability space (LEPS)
- correlation coefficient
- anomaly correlation
- S1 score

**Advantages:** Simple, familiar, long historical records

**Disadvantage:** Not very revealing as to what's going wrong in the forecast



## Mean error (bias)

$$\text{Mean Error} = \frac{1}{N} \sum_{i=1}^N (F_i - O_i)$$

Measures: Average difference between forecast and observed values

Verifies this attribute?

Location

Size

Shape

Mean value



Maximum value

Spatial variability



## Mean absolute error

$$MAE = \frac{1}{N} \sum_{i=1}^N |F_i - O_i|$$

Measures: Average magnitude of forecast error

## Root mean square error

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (F_i - O_i)^2}$$

Measures: Error magnitude, with large errors having a greater impact than in the MAE

Verifies this attribute?

Location

Size

Shape

Mean value ✓

Maximum value ✓

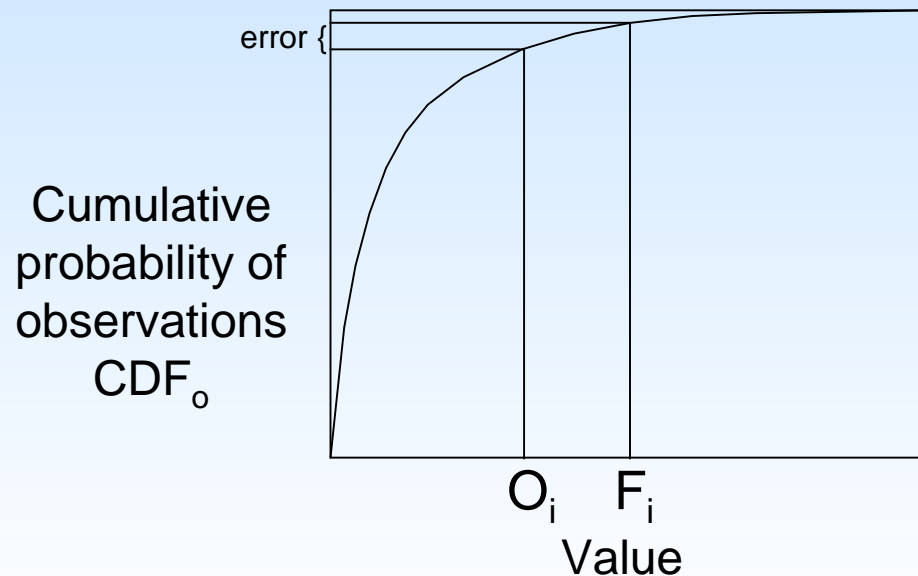
Spatial variability



## Linear error in probability space (LEPS)

$$LEPS = \frac{1}{N} \sum_{i=1}^N |CDF_o(F_i) - CDF_o(O_i)|$$

Measures: Probability error - does not penalize going out on a limb when it is justified.



Verifies this attribute?

Location

Size

Shape

Mean value ✓

Maximum value ✓

Spatial variability



## Correlation coefficient

$$r = \frac{\sum (F - \bar{F})(O - \bar{O})}{\sqrt{\sum (F - \bar{F})^2} \sqrt{\sum (O - \bar{O})^2}}$$

Measures: Spatial correspondence between forecast pattern and observed pattern

## Anomaly correlation

$$AC = \frac{\sum (F - C)(O - C)}{\sqrt{\sum (F - C)^2} \sqrt{\sum (O - C)^2}}$$

Measures: Correlation of forecast and observed anomalies relative to climatology  $C$  at each gridpoint

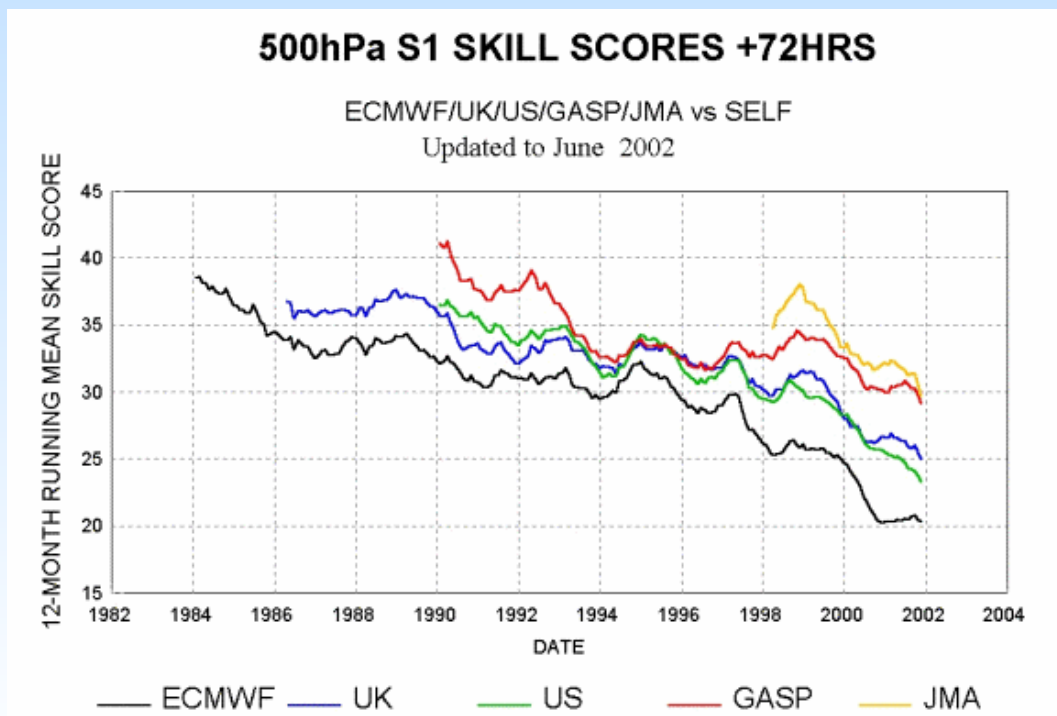
<u>Verifies this attribute?</u>	
Location	✓
Size	
Shape	✓
Mean value	
Maximum value	
Spatial variability	✓



# S1 score

$$S1 = \frac{\sum_{adjacent\ pairs} |\Delta F - \Delta O|}{\sum_{adjacent\ pairs} \max(|\Delta F|, |\Delta O|)} \times 100$$

Measures: Accuracy of forecasts of gradients



<u>Verifies this attribute?</u>	
Location	✓
Size	
Shape	✓
Mean value	
Maximum value	
Spatial variability	✓

# Categorical statistics

Measure the correspondence between forecast and observed *occurrence of events* at gridpoints

Examples:

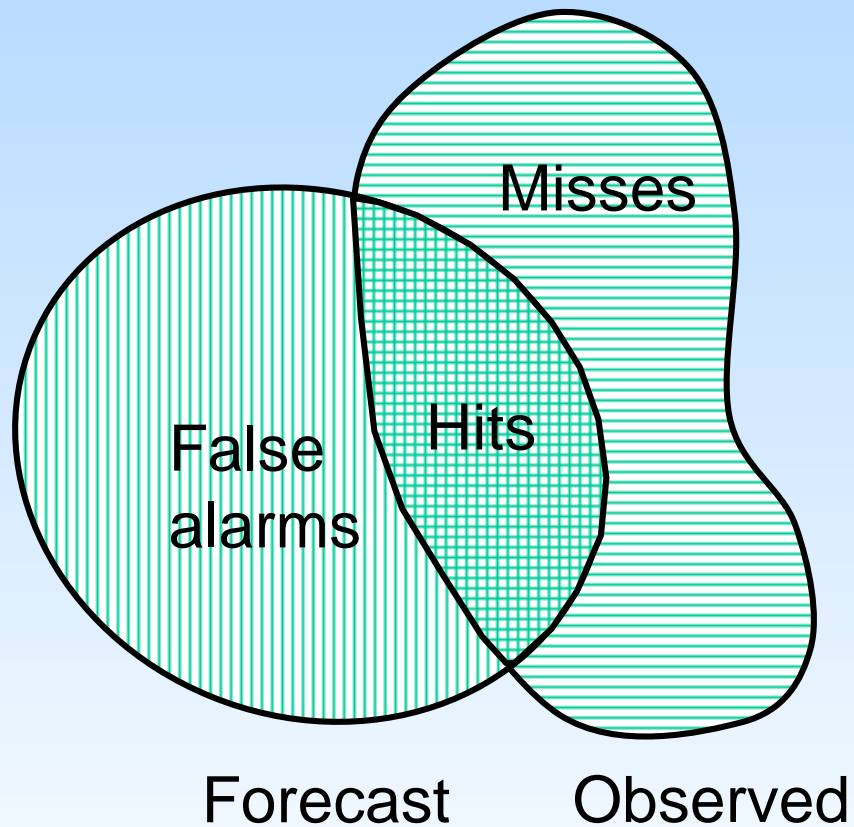
- bias score
- probability of detection
- false alarm ratio
- threat score
- equitable threat score
- odds ratio
- Hanssen and Kuipers score
- relative operating characteristic
- Heidke skill score

**Advantages:** Simple, familiar, ~long historical records

**Disadvantage:** Not very revealing



# Categorical statistics



		Predicted	
		yes	no
Observed	yes	hits	misses
	no	false alarms	correct negatives



## Bias score

$$BIAS = \frac{hits + false\ alarms}{hits + misses}$$

Measures: Ratio of forecast area to observed area

Verifies this attribute?

Location

Size



Shape

Mean value

Maximum value

Spatial variability



Probability of Detection

$$POD = \frac{hits}{hits + misses}$$

False Alarm Ratio

$$FAR = \frac{false\ alarms}{hits + false\ alarms}$$

Threat score (critical success index)

$$TS = CSI = \frac{hits}{hits + misses + false\ alarms}$$

Equitable threat score

$$ETS = \frac{hits - hits_{random}}{hits + misses + false\ alarms - hits_{random}}$$

Odds ratio

$$OR = \frac{hits * correct\ negatives}{misses * false\ alarms}$$

Verifies this attribute?

Location ✓

Size ✓

Shape ✓

Mean value

Maximum value

Spatial variability



## Hanssen and Kuipers discriminant (true skill statistic)

$$HK = \frac{\text{hits}}{\text{hits} + \text{misses}} - \frac{\text{false alarms}}{\text{false alarms} + \text{correct negatives}}$$

Measures: Ability of the forecast to separate the "yes" cases from the "no" cases.

## Relative operating characteristic

$$ROCArea = \frac{1}{2} \left[ \frac{\text{hits}}{\text{hits} + \text{misses}} - \frac{\text{false alarms}}{\text{false alarms} + \text{correct negatives}} + 1 \right]$$
$$= \frac{1}{2} (HK + 1)$$

Measures: Ability of the forecast to separate the "yes" cases from the "no" cases

Verifies this attribute?

Location ✓

Size ✓

Shape ✓

Mean value

Maximum value

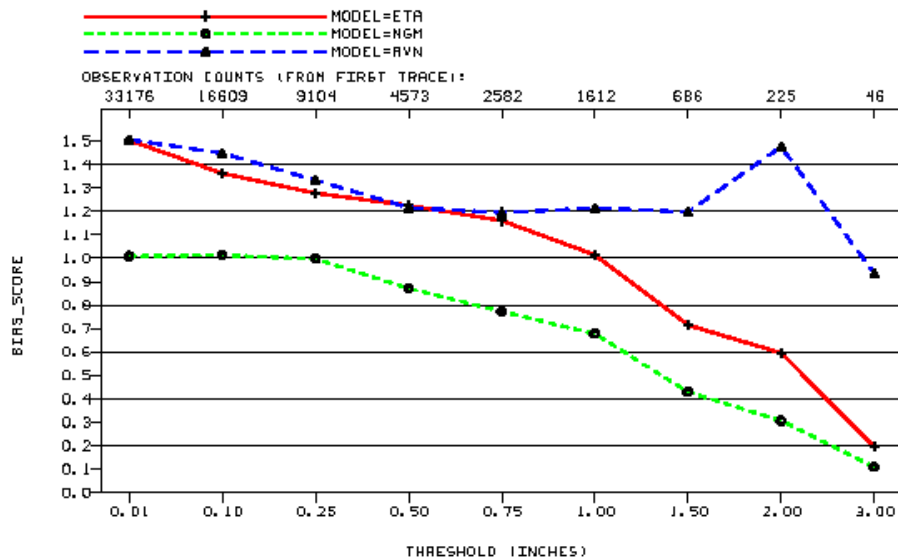
Spatial variability



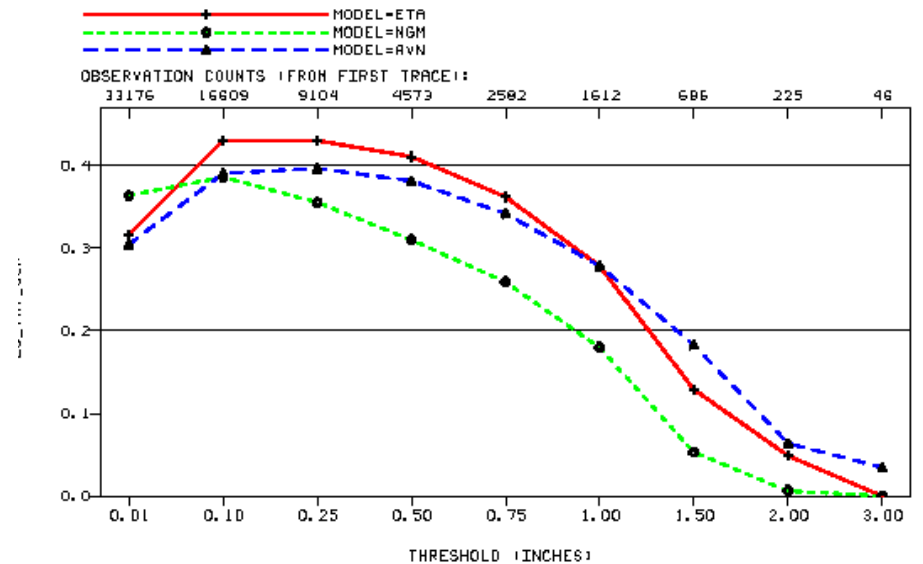
# NCEP verification of QPFs from NWP models

STAT=FHO PARAM=APCP/24 FHOUR=24+36+48 V\_ANL=MB\_PCP V\_RGN=6211/RFC LEVEL=5FC  
 VYMDH=200103010000-200103312300

STAT=FHO PARAM=APCP/24 FHOUR=24+36+48 V\_ANL=MB\_PCP V\_RGN=6211/RFC LEVEL=5FC  
 VYMDH=200103010000-200103312300



Bias score



Equitable threat score

Rain threshold varies from light to heavy

# Distributions oriented view

		Observations											N	p(f)
		≤ -25	-20	-15	-10	-5	0	5	10	15	20	≥ 25		
		a)												
Forecasts	≤ -25	5	1	0	0	0	0	0	0	0	0	0	6	1.0
	-20	3	6	0	0	0	0	0	0	0	0	0	9	1.5
	-15	4	4	9	3	0	0	0	0	0	0	0	20	3.4
	-10	2	3	9	18	6	4	0	0	0	0	0	42	7.1
	-5	0	1	2	15	36	21	3	1	0	0	0	79	13.4
	0	0	0	1	7	29	122	49	8	0	0	0	216	36.6
	5	0	1	1	0	3	40	61	26	2	1	0	135	22.9
	10	0	0	0	0	0	3	18	28	10	1	0	60	10.2
	15	0	1	0	0	0	0	1	2	10	2	2	17	2.9
	20	0	0	0	0	0	0	0	0	0	3	2	5	0.8
	≥ 25	0	0	0	0	0	0	0	0	0	1	0	1	0.2
N	14	16	22	43	74	190	132	65	22	8	4	298		
p(x)	2.4	2.7	3.7	7.3	12.5	32.2	22.4	11.0	3.7	1.4	0.7			

**Advantage:** Much more instructive picture of forecast performance

**Disadvantage:** Lots of numbers

Verifies this attribute?

Location ✓

Size ✓

Shape ✓

Mean value ✓

Maximum value ✓

Spatial variability

## Heidke skill score (K distinct categories)

$$HSS = \frac{\sum_{k=1}^K P(F_k, O_k) - \sum_{k=1}^K P(F_k)P(O_k)}{1 - \sum_{k=1}^K P(F_k)P(O_k)}$$

Measures: Skill of the forecast in predicting the correct category, relative to that of random chance

<u>Verifies this attribute?</u>	
Location	✓
Size	✓
Shape	✓
Mean value	✓
Maximum value	✓
Spatial variability	

# Scale decomposition methods

Measure the correspondence between the forecasts and observations at *different spatial scales*

Examples:

- 2D Fourier decomposition
- wavelet decomposition
- upscaling

**Advantages:** Scales on which largest errors occur can be isolated, can "clean up" noisy data, forecasts and observations can be on different grids

**Disadvantages:** Less intuitive, can be mathematically tricky



# Discrete wavelet transforms

Briggs and Levine (1997)

Concept: Decompose fields into scales representing different detail levels. Test whether the forecast resembles the observations at each scale.

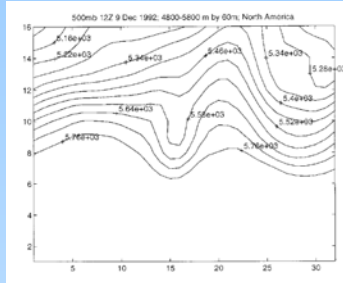
Method: Discrete wavelet transformation

Measures, for each scale:

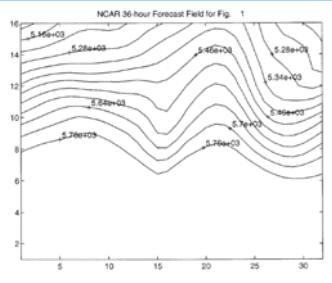
- % anomaly correlation
- % MSE
- linear correlation
- RMSE
- energy ratio

<u>Verifies this attribute?</u>	
Location	✓
Size	✓
Shape	✓
Mean value	✓
Maximum value	
Spatial variability	✓

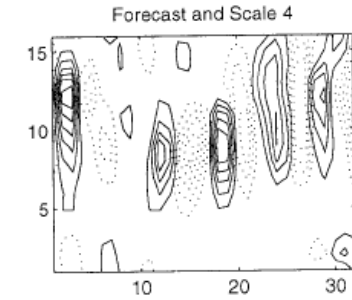
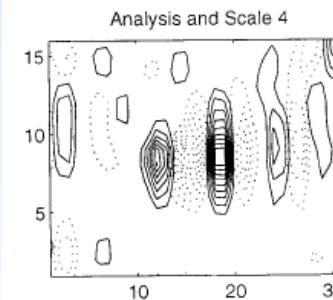
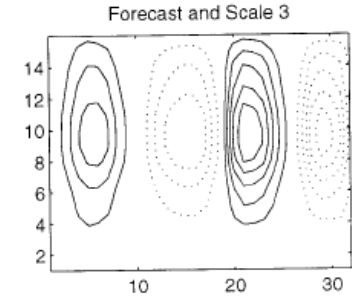
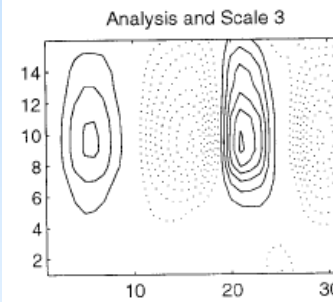
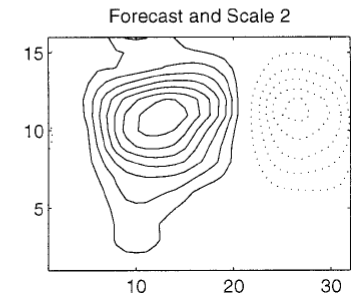
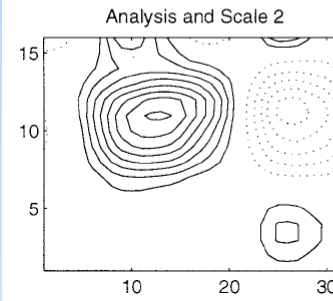
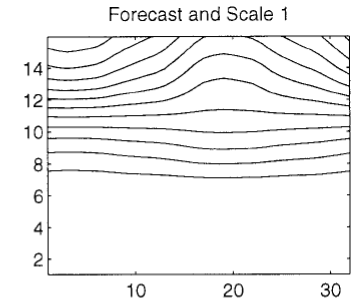
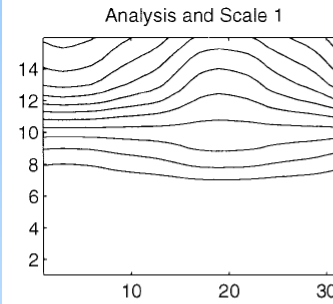




analysis



forecast



Scale	1	2	3	4	5
$\%ACC_h$	0.80	0.09	0.10	0.01	0.00
$\%MSE_h$	0.48	0.16	0.20	0.08	0.08
$r_h$	0.998	0.968	0.978	0.808	0.409
$rmse_h$	11.56	6.98	7.87	4.85	5.68
$ER_h$	0	0.01	0.25	0.03	0.51

etc.



# Multiscale statistical organization

Zepeda-Arce et al. (2000)

Concept: Observed precipitation patterns have multi-scale spatial and spatio-temporal organization. Test whether the forecast reproduces this organization.

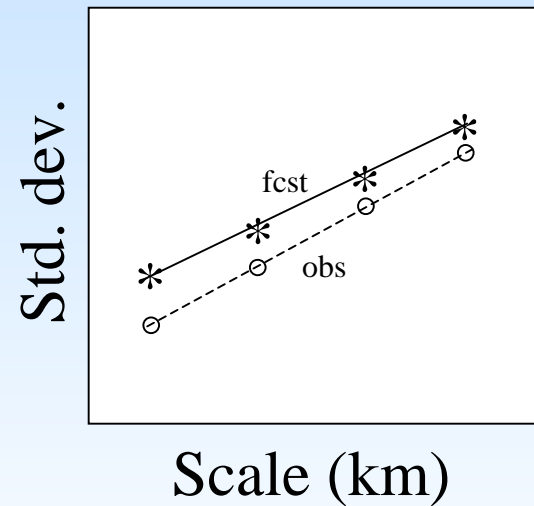
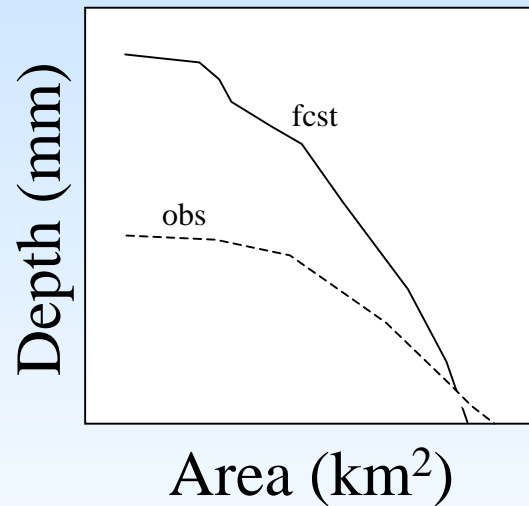
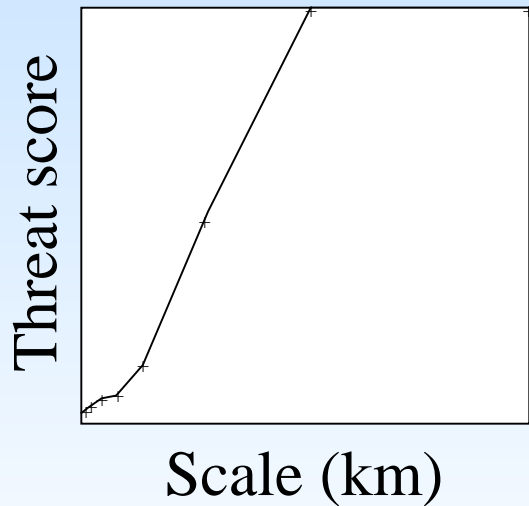
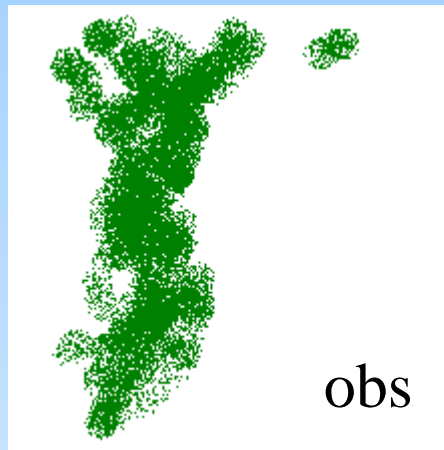
Method: Start with fine scale, average to coarser scale

Measures:

- TS vs. scale
- depth vs. area
- spatial scaling parameter
- dynamic scaling exponent

<u>Verifies this attribute?</u>	
Location	✓
Size	✓
Shape	✓
Mean value	✓
Maximum value	
Spatial variability	✓





# Entity-based methods

Use pattern matching to associate forecast and observed entities ("blobs"). Verify the properties of the entities.

Examples:

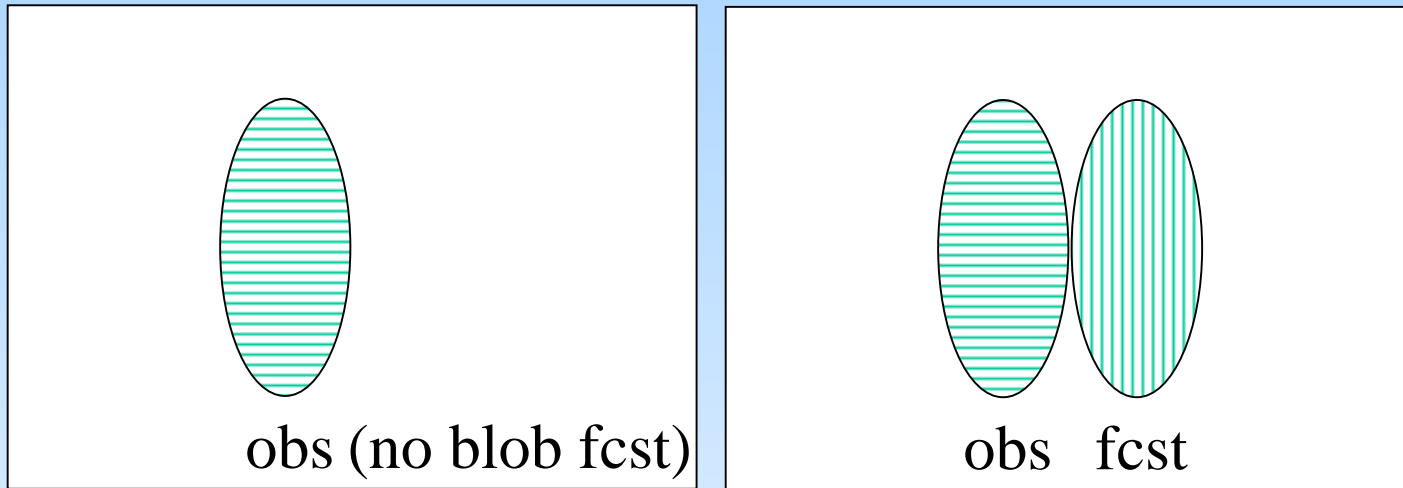
- CRA (contiguous rain area) verification
- object-based diagnostic approach
- convergence line verification

**Advantages:** Intuitive, measures location errors, allows error decomposition, addresses "double penalty"

**Disadvantage:** May fail if forecast does not sufficiently resemble observations



# "Double penalty"



This scores better than this according to most continuous and categorical statistics!

# CRA (entity) verification

Ebert and McBride (2000)

Concept: Verify the properties of the forecast entities against observed entities

Method: Pattern matching to determine location error, error decomposition, event verification

Measures:

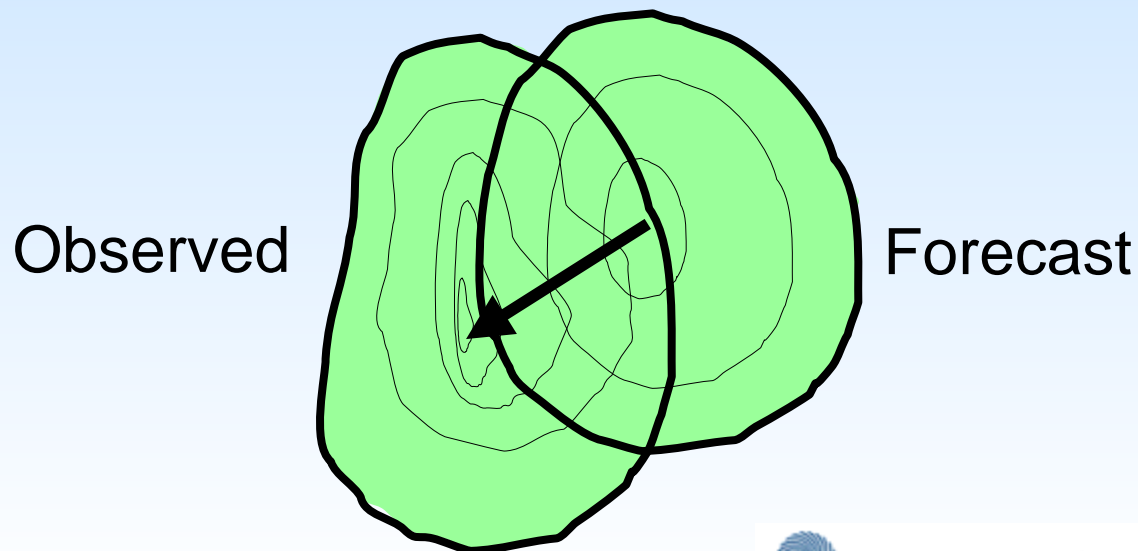
- location error
- size error
- error in mean, max values
- pattern error

<u>Verifies this attribute?</u>	
Location	✓
Size	✓
Shape	✓
Mean value	✓
Maximum value	✓
Spatial variability	✓



Determine the location error using *pattern matching*:

- Horizontally translate the QPF until the total squared error between the forecast and the observations is minimized in the shaded region. Other possibilities: maximum correlation, maximum overlap
- The displacement is the vector difference between the original and final locations of the forecast.



## CRA error decomposition

The total mean squared error (MSE) can be written as:

$$MSE_{total} = MSE_{displacement} + MSE_{volume} + MSE_{pattern}$$

The difference between the mean square error before and after translation is the contribution to total error due to *displacement*,

$$MSE_{displacement} = MSE_{total} - MSE_{shifted}$$

The error component due to *volume* represents the bias in mean intensity,

$$MSE_{volume} = (\bar{F} - \bar{X})^2$$

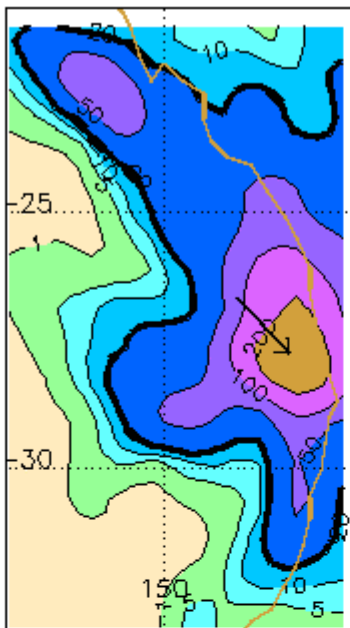
where  $\bar{F}$  and  $\bar{X}$  are the CRA mean forecast and observed values after the shift.

The *pattern error* accounts for differences in the fine structure of the forecast and observed fields,

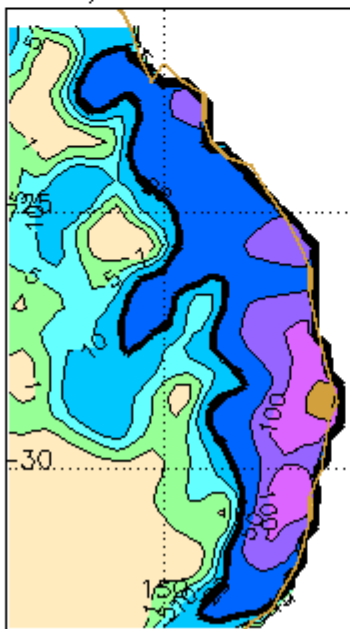
$$MSE_{pattern} = MSE_{shifted} - MSE_{volume}$$



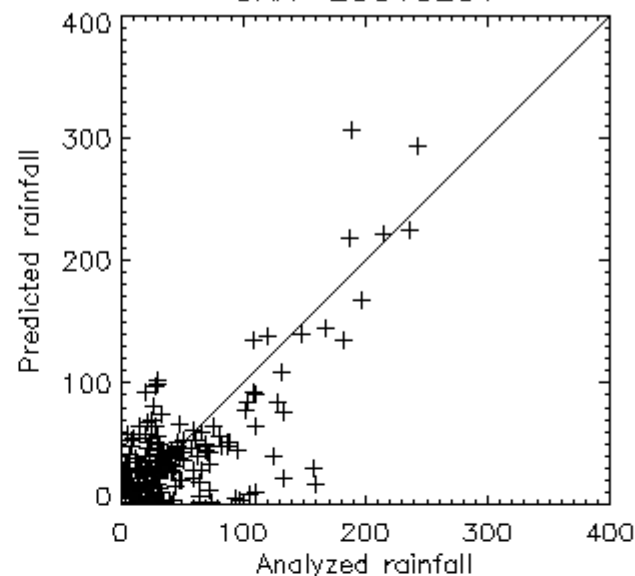
LAPS375 fcst 20010201



Analysis 20010201



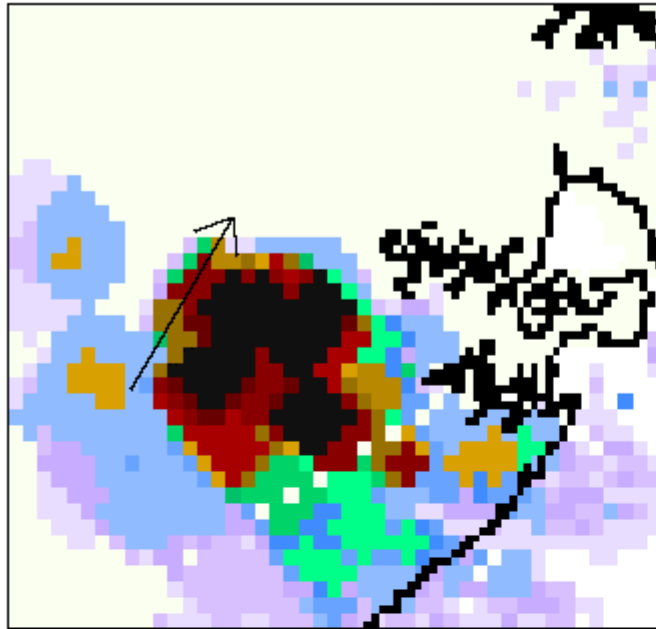
CRA 20010201



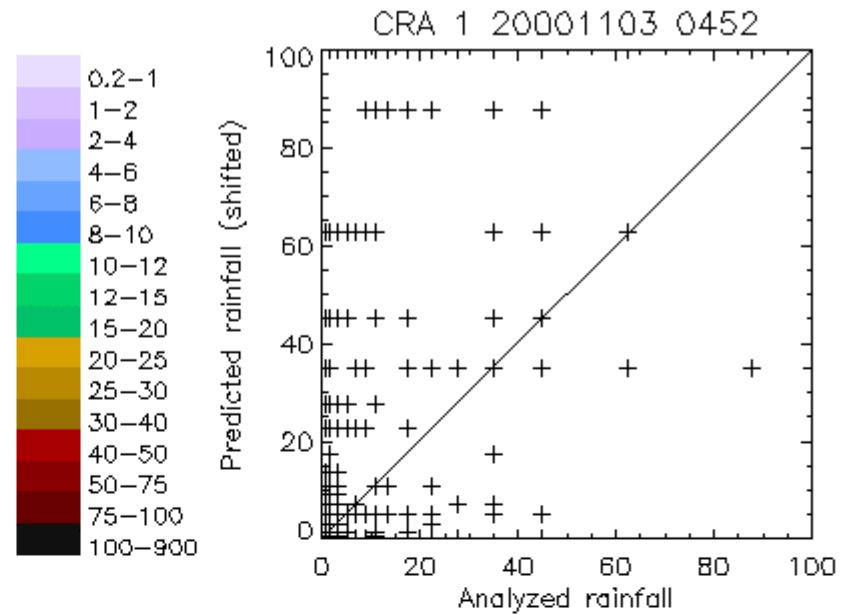
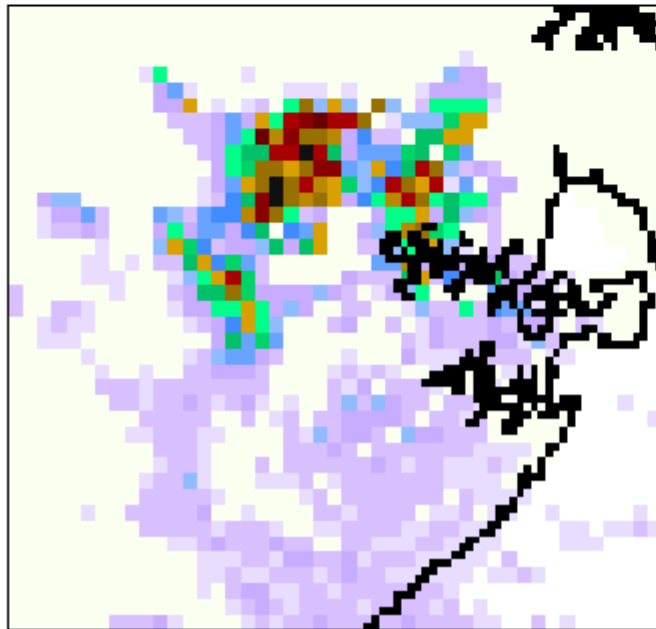
LAPS375 00-24 fcst 20010201 n=232  
 (-32.75°,147.12°) to (-21.50°,153.50°)  
 Verif. grid=0.375° CRA threshold=20.0 mm/d

	Analysed	Forecast
# gridpoints $\geq 20$ mm/d	159	192
Average rainrate (mm/d)	61.86	45.27
Maximum rain (mm/d)	242.90	306.91
Rain volume (km <sup>3</sup> )	15.17	13.40
Displacement (E,N) = [-1.12°,1.12°]		
	Original	Shifted
RMS error (mm/d)	60.48	32.65
Correlation coefficient	0.469	0.731
Error Decomposition:		
Displacement error	70.9%	
Volume error	0.7%	
Pattern error	28.5%	

AutoNowcaster 30 min fcst valid 20001103 0452

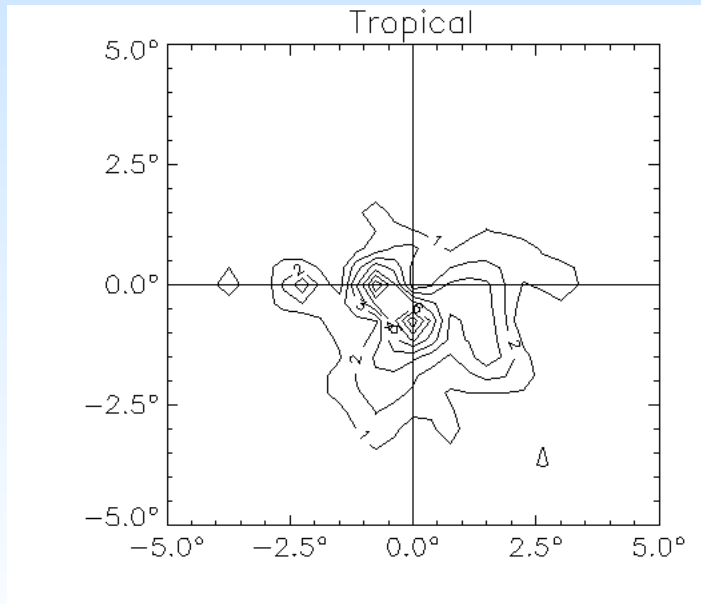
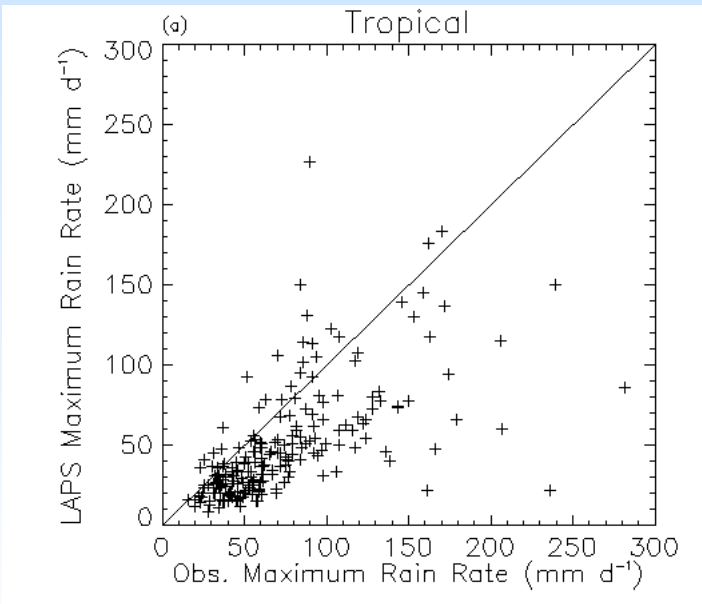
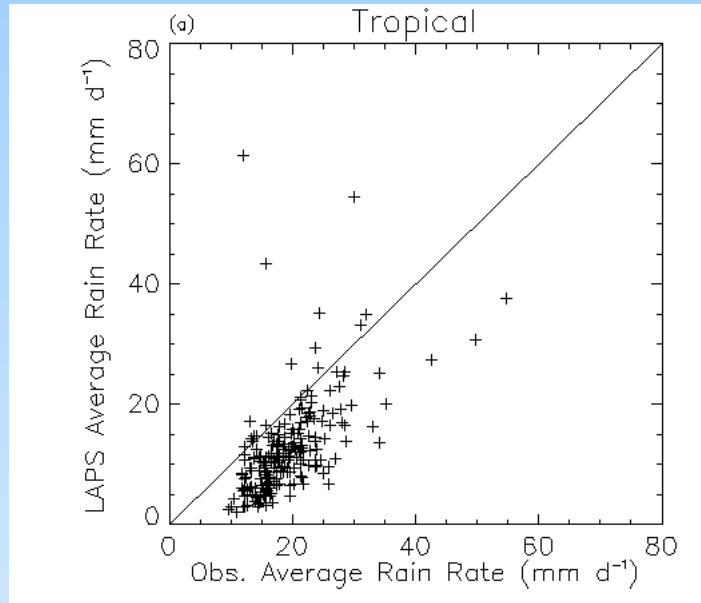
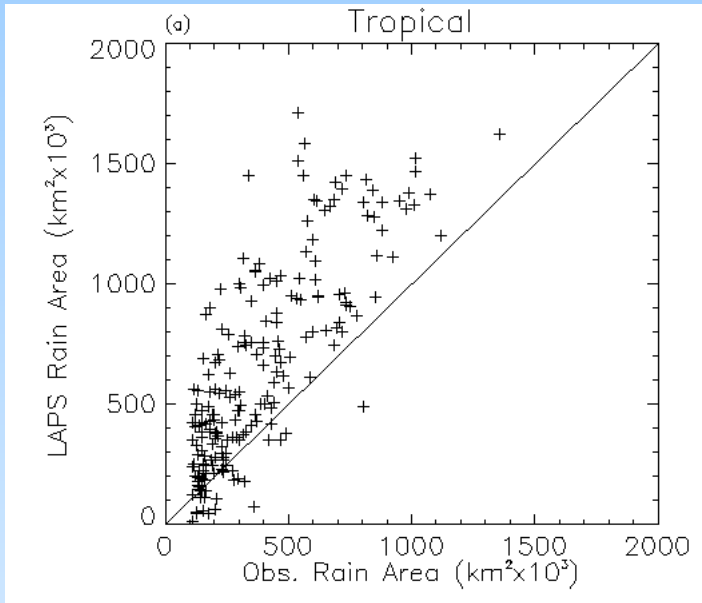


Radar rain rate valid 20001103 0452

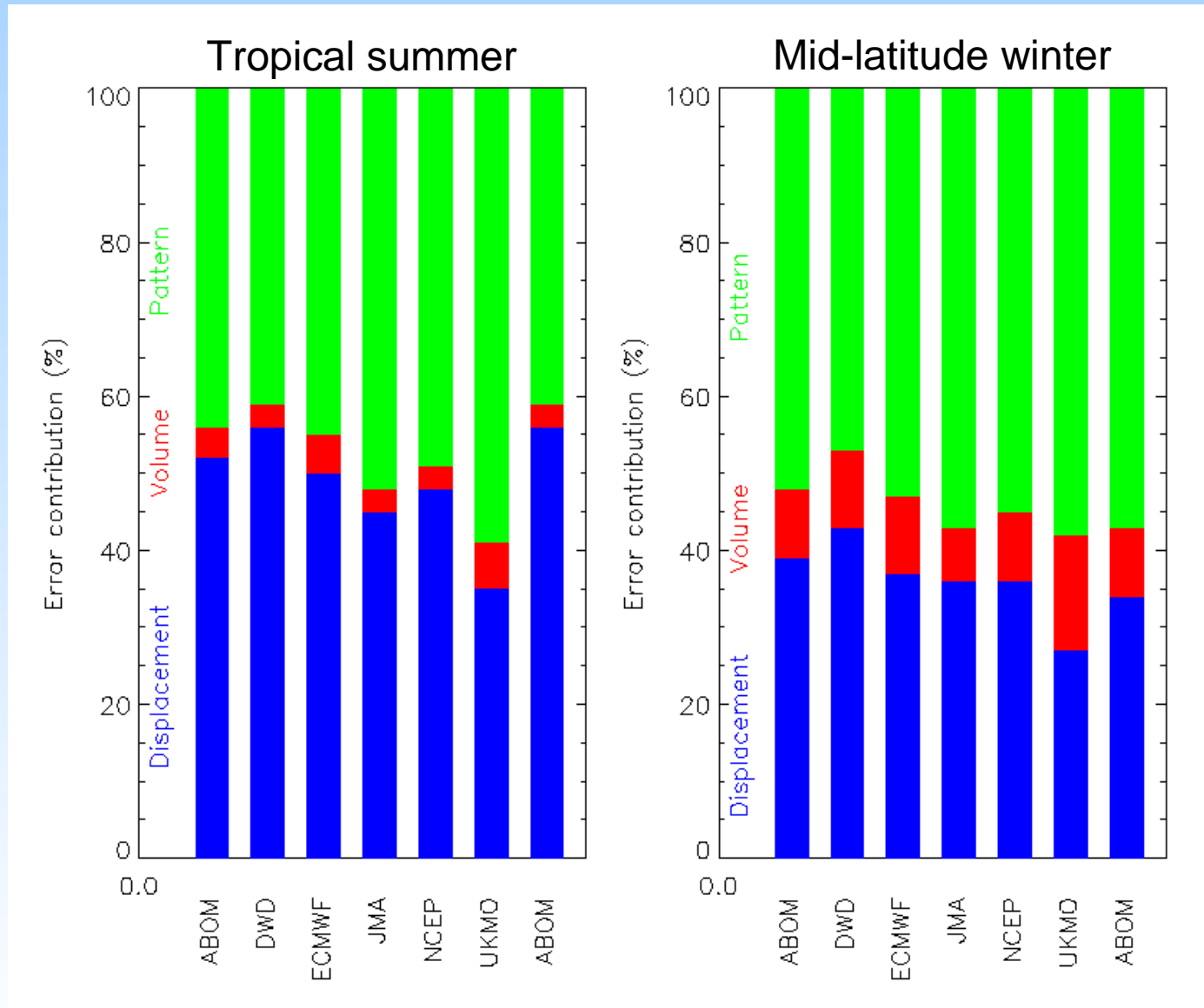


AutoNowcaster 0030min fcst 20001103 0452 n=945  
 (-34.20°,150.76°) to (-33.91°,151.16°)  
 Verif. grid=1 km CRA threshold=5 mm/hr

	Analysed	Forecast
# gridpoints $\geq$ 5 mm/hr	201	566
Average rainrate (mm/hr)	23.00	25.81
Maximum rain (mm/hr)	100.00	100.00
Rain volume ( $\text{km}^3 \cdot 10^{-3}$ )	4.62	14.60
Displacement (E,N) = [-7 km,-11 km]		
	Original	Shifted
RMS error (mm/hr)	35.98	26.44
Correlation coefficient	-0.093	0.518
Error Decomposition:		
Displacement error	46.0%	
Volume error	8.6%	
Pattern error	45.4%	



# Error decomposition for 24 h QPFs over Australia



# Object-based diagnostic approach

Brown et al. (2002)

Concept: Characterize forecast and observed precipitation/convective regions in a natural way as geometric objects

Method: Direct comparison of fundamental attributes of the forecasts and observations

Measures:

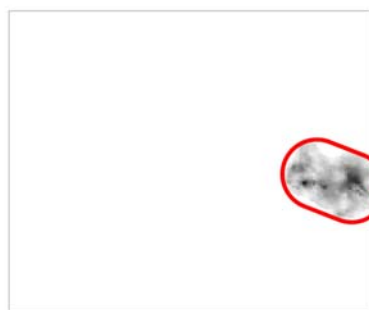
- location error
- shape and orientation error
- size error
- error in mean, max values

<u>Verifies this attribute?</u>	
Location	✓
Size	✓
Shape	✓
Mean value	✓
Maximum value	✓
Spatial variability	✓

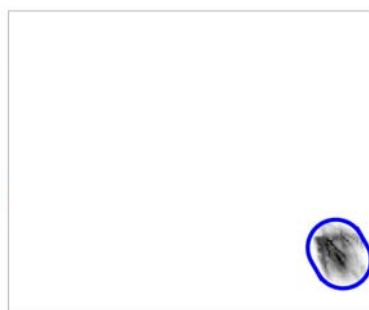


# Example: Precip forecasts and observations

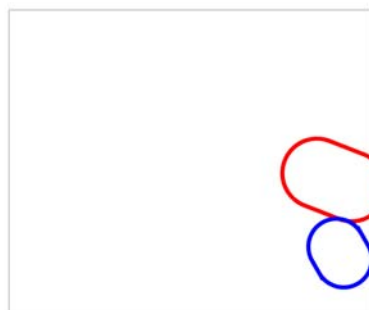
## Shapes/objects and matching:



Stage 4 data  
Area 1403



WRF4 data  
Area 809

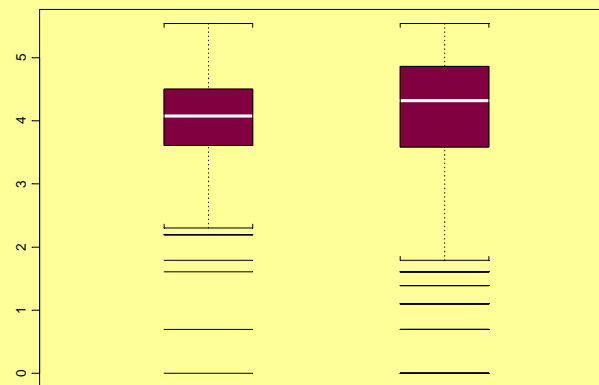


Int Ratio 0.14 %  
SD Ratio 157.38 %



Int Ratio 57.59 %  
SD Ratio 42.41 %

## Precipitation intensities:



Stage 4

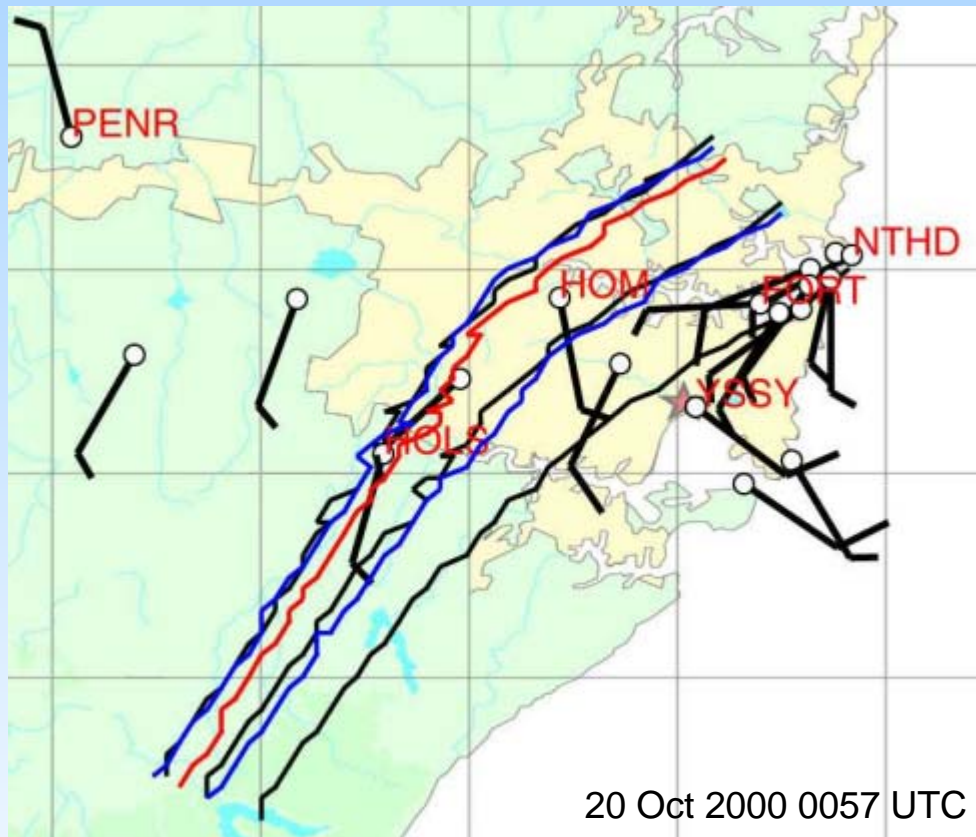
WRF



# Line verification

P. Nurmi (2002)

Determine mean position error by measuring average distance from each vertex in forecast line to nearest point in observed line.



— Observed  
— 30 min fcst  
— 60 min fcst

Verifies this attribute?	
Location	✓
Size	✓
Shape	✓
Mean value	
Maximum value	
Spatial variability	

# Event-oriented methods

Verify the bulk properties of defined events

Examples:

- mesoscale classification
- compositing
- event verification

**Advantages:** Intuitive, verifies situations rather than point values

**Disadvantage:** Require subjective input in choosing classes / definitions of events



# Mesoscale classification

Baldwin et al. (2001)

Concept: Spatial characteristics can be objectively classified and have meteorological significance

Method: Use pattern recognition techniques to classify small domains into mesoscale types (events), examine joint distribution of forecast and observed events



1111221111  
 1122221441  
 1222333444  
 2233344555  
 3311455511

observed class

1112221111  
 1222221331  
 1222233344  
 2223444552  
 2344425522

forecast class

		Forecast class					N
		1	2	3	4	5	
Observed class	1	11	4	0	2	0	17
	2	0	11	0	0	0	11
	3	0	3	4	1	0	8
	4	0	0	3	5	0	8
	5	0	2	0	0	4	6
N		11	20	7	8	4	50

<u>Verifies this attribute?</u>	
Location	✓
Size	
Shape	✓
Mean value	✓
Maximum value	✓
Spatial variability	✓

# Nachamkin's compositing method

Concept: Composite techniques allow bulk properties of events to be estimated and compared.

Method: Isolate meteorological events using a rules-based algorithm and composite on a relative grid centered on each event. Compare bulk properties of forecast with bulk properties of observations.

<u>Verifies this attribute?</u>	
Location	✓
Size	✓
Shape	✓
Mean value	✓
Maximum value	✓
Spatial variability	✓



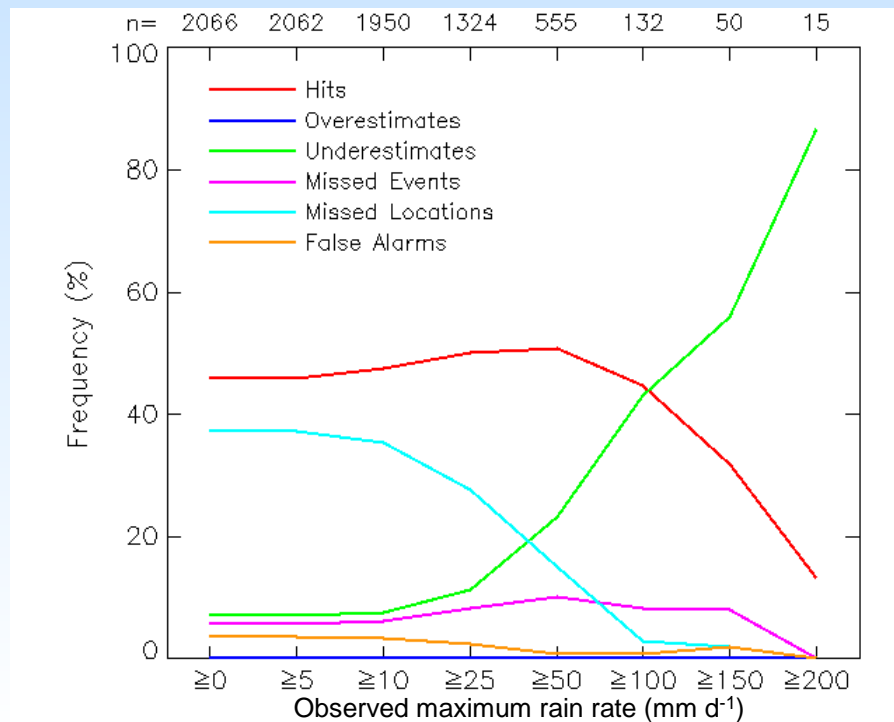
# Event verification

Ebert and McBride (2000)

Displacement  
of forecast  
pattern

Forecast Maximum Value

		Forecast Maximum Value		
		Too Little	Approx. Correct	Too Much
Displacement of forecast pattern	Close	<i>Under-estimate</i>	<i>Hit</i>	<i>Over-estimate</i>
	Far	<i>Missed Event</i>	<i>Missed Location</i>	<i>False Alarm</i>



<u>Verifies this attribute?</u>	
Location	✓
Size	
Shape	
Mean value	
Maximum value	✓
Spatial variability	

# Which methods verify which attributes?

	Visual ("eyeball")	Contin- uous statistics	Cate- gorical statistics	Joint distrib- ution	Scale decom- position	Entity- based	Event- oriented
Location	✓	✓	✓	✓	✓	✓	✓
Size	✓		✓	✓	✓	✓	✓
Shape	✓	✓	✓	✓	✓	✓	✓
Mean value	✓	✓		✓	✓	✓	✓
Maximum value	✓	✓		✓		✓	✓
Spatial variability	✓	✓			✓	✓	✓

# Conclusions

- The most effective spatial verification method is still visual ("eyeball") verification
- Categorical statistics based on yes-no discrimination are probably the least informative of all of the verification methods
- The newer "scientific" verification methods (scale decomposition, entity-based, and event-oriented) give a more complete picture of spatial forecast performance



# References

Baldwin et al. (2001)

Briggs, W.M. and R.A. Levine, 1997: Wavelets and field forecast verification. *Mon. Wea. Rev.*, **125**, 1329-1341.

Brown, B.G., R. Bullock, C. Davis, K. Manning, R. Morss, and C. Mueller, 2002: An object-based diagnostic approach for QPF and convective forecast verification. Workshop on Making Verification More Meaningful, Boulder, CO, 30 July - 1 August 2002.

Ebert, E.E. and J.L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrology*, **239**, 179-202.

Zepeda-Arce et al. (2000)

