Methods for verifying spatial forecasts

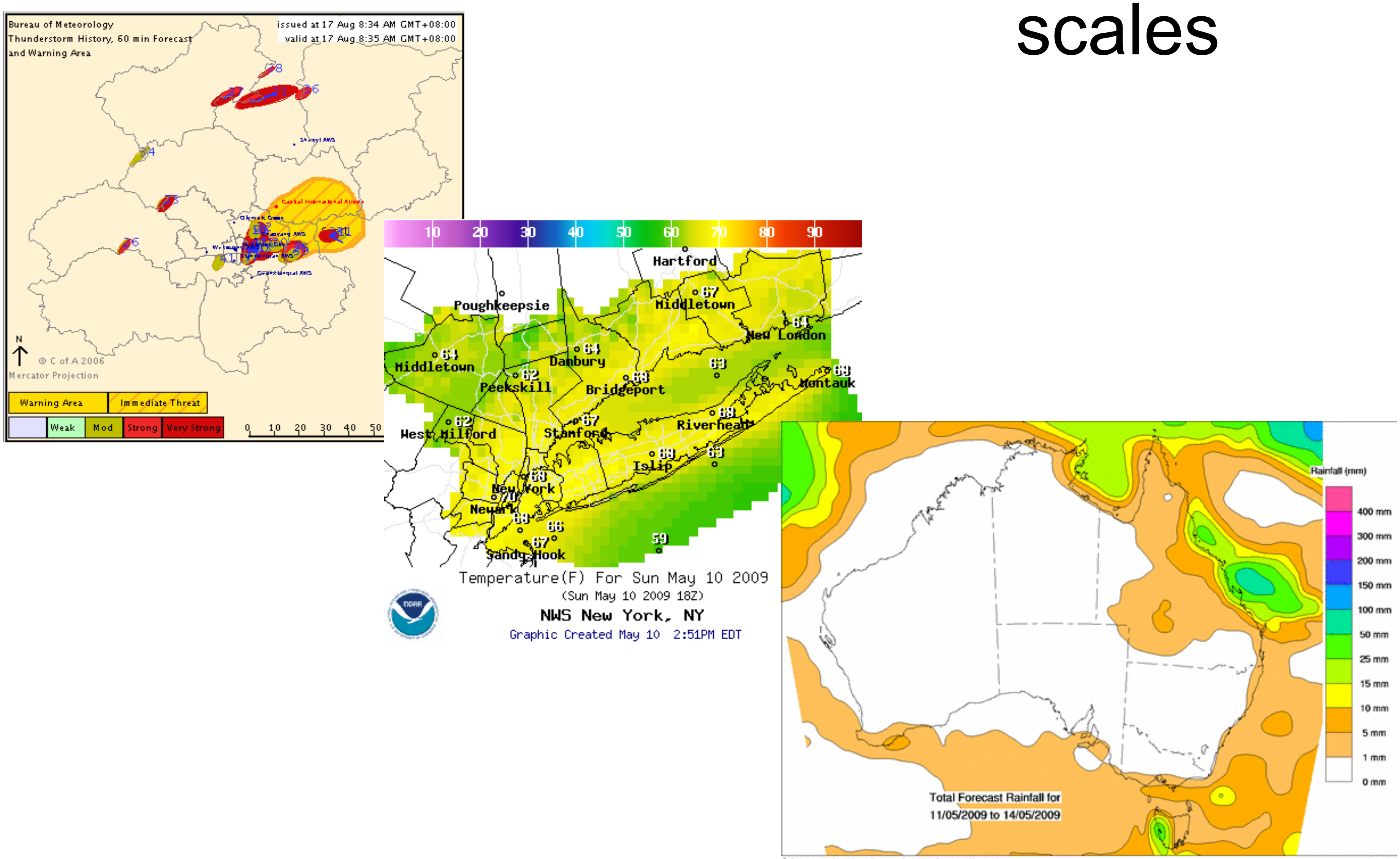
Beth Ebert

Centre for Australian Weather and Climate Research (CAWCR)
Bureau of Meteorology, Melbourne, Australia

Acknowledgements: Barb Brown, Barbara Casati, Marion Mittermaier

4th Int'l Verification Methods Workshop, Helsinki, 4-6 June 2009

Spatial forecasts are made at many scales

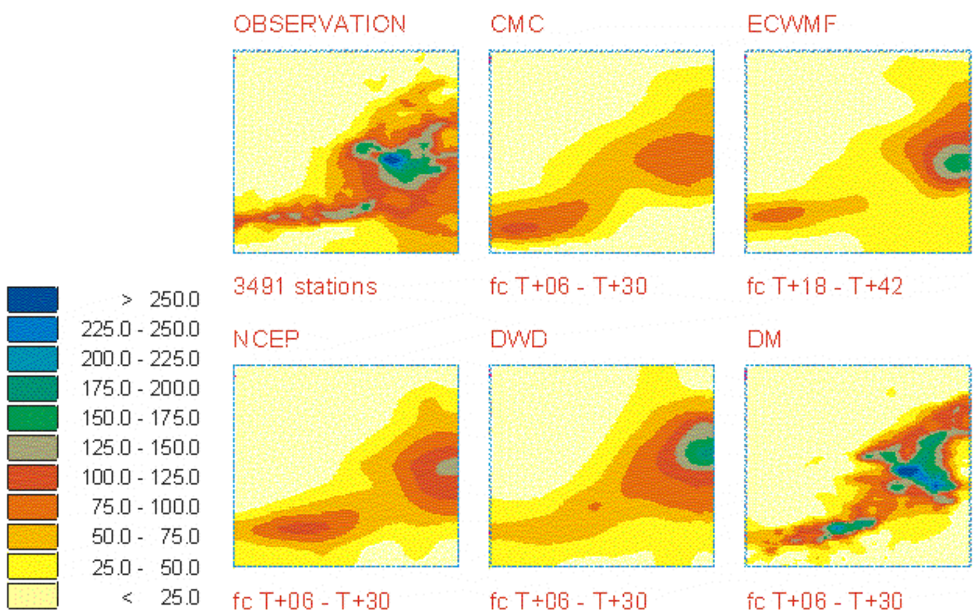


Visual ("eyeball") verification

Visually compare maps of forecast and observations

Advantage: "A picture tells a thousand words…"

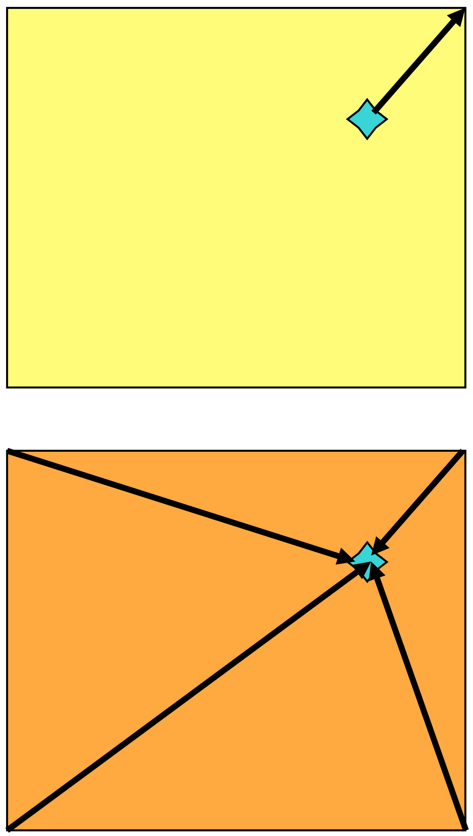
Disadvantages: Labor intensive, not quantitative, subjective



Matching forecasts and observations

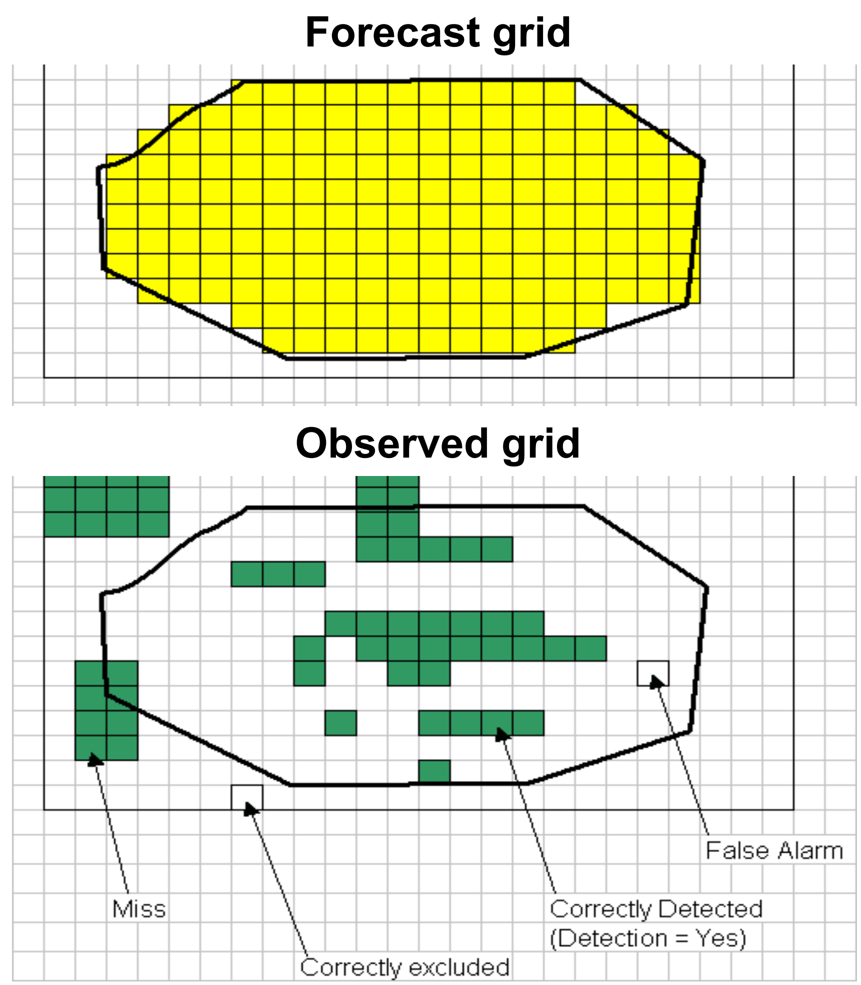
- Point-to-grid and
  grid-to-point

- Matching approach can impact the results of the verification



Matching forecasts and observations

- Grid to grid approach
  - Overlay forecast and observed grids
  - Match each forecast and observation

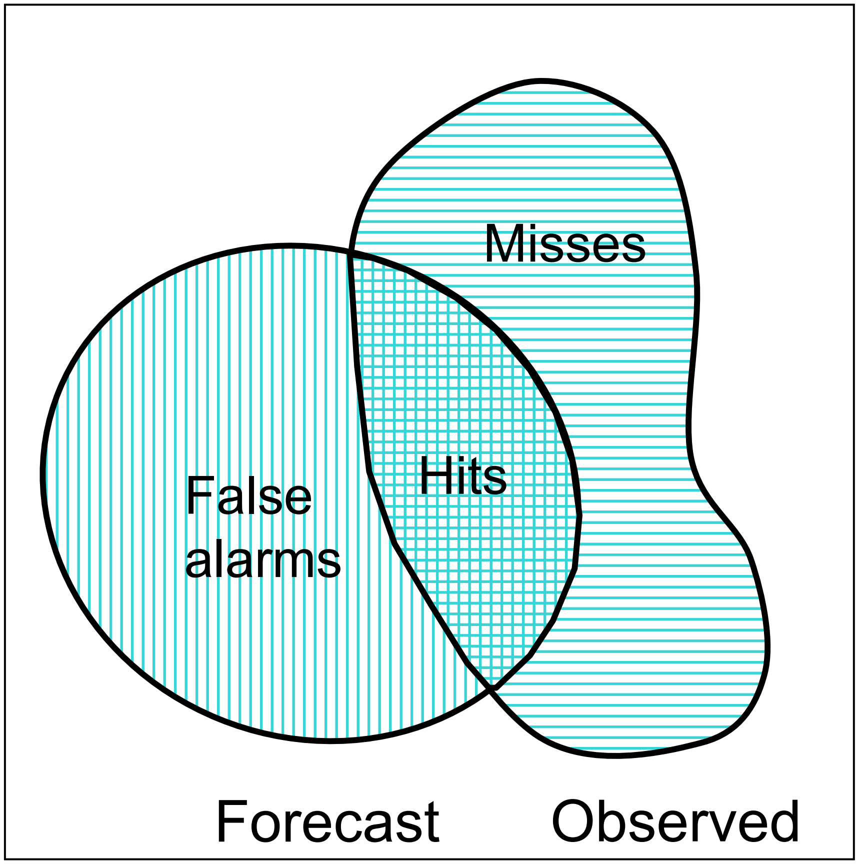


Traditional verification approaches

Compute statistics on forecast-observation pairs

- Continuous values (e.g., precipitation amount, temperature, NWP variables):
  - mean error, MSE, RMSE, correlation
  - anomaly correlation, S1 score
- Categorical values (e.g., precipitation occurrence):
  - Contingency table statistics (POD, FAR, Heidke skill score, equitable threat score, Hanssen-Kuipers statistic…)

Traditional spatial verification using categorical scores



Contingency Table

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

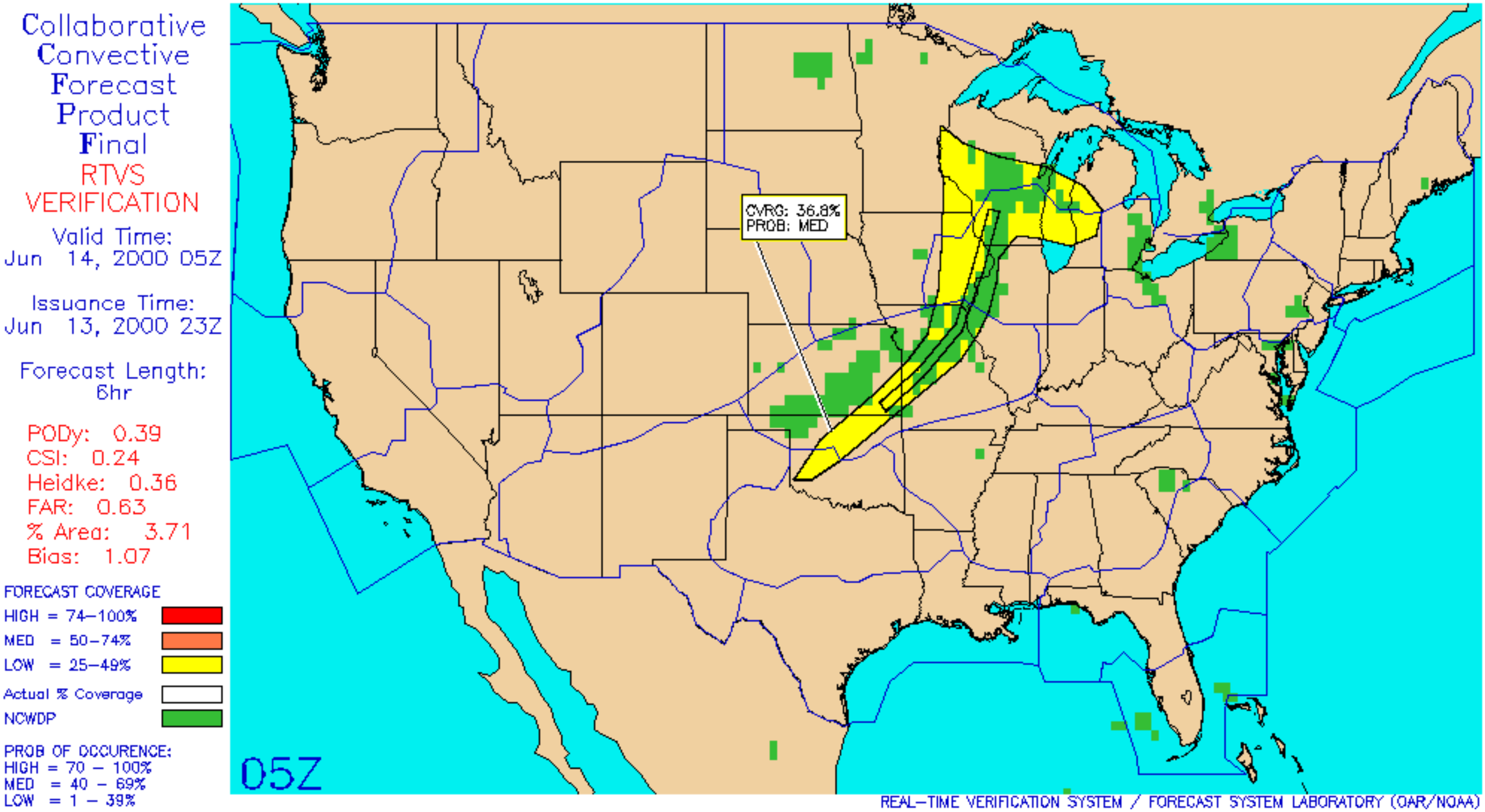
$$FBI = \frac{hits + false\ alarms}{hits + misses}$$

$$POD = \frac{hits}{hits + misses} \qquad FAR = \frac{false\,alarms}{hits + false\,alarms}$$

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

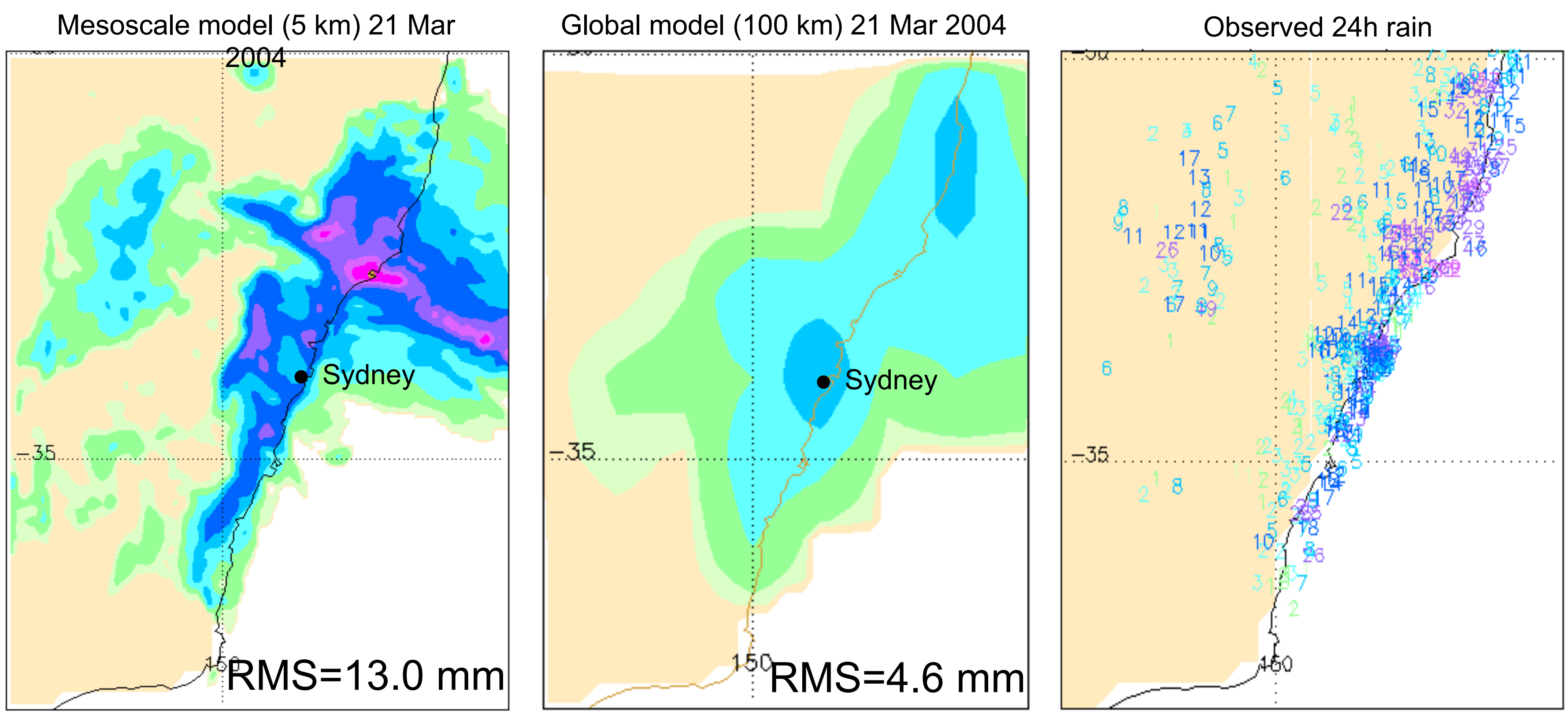
$$ETS = \frac{hits - hits_{random}}{hits + misses + false\,alarms - hits_{random}}$$

PODy=0.39, FAR=0.63, CSI=0.24



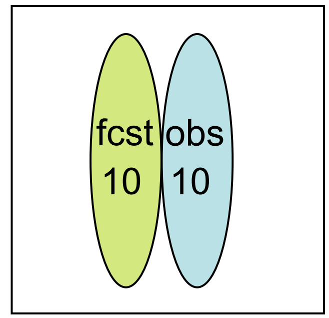
High vs. low resolution

Which forecast would you rather use?

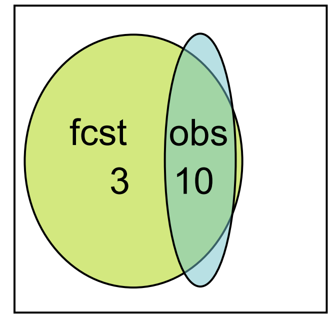


Traditional spatial verification

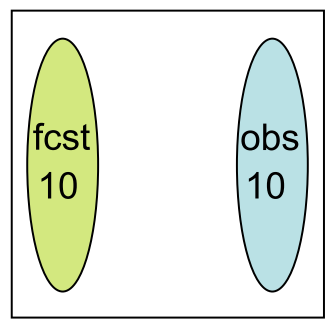
- Requires an exact match between forecasts and observations at every grid point
  - Problem of "double penalty" - event predicted where it did not occur, no event predicted where it did occur
- Traditional scores do not say very much about the source or nature of the errors



Hi res forecast
RMS ~ 4.7
POD=0, FAR=1
TS=0



Low res forecast
RMS ~ 2.7
POD~1, FAR~0.7
TS~0.3



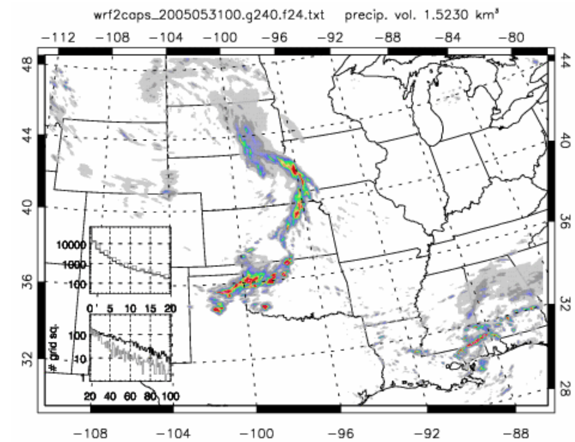
What's missing?

- Traditional approaches provide overall measures of skill but…
- They provide minimal ***diagnostic*** information about the forecast:
  - What went wrong? What went right?
  - Does the forecast look realistic?
  - How can I improve this forecast?
  - How can I use it to make a decision?
- Best performance for ***smooth*** forecasts
- Some scores are insensitive to the ***size*** of the errors…

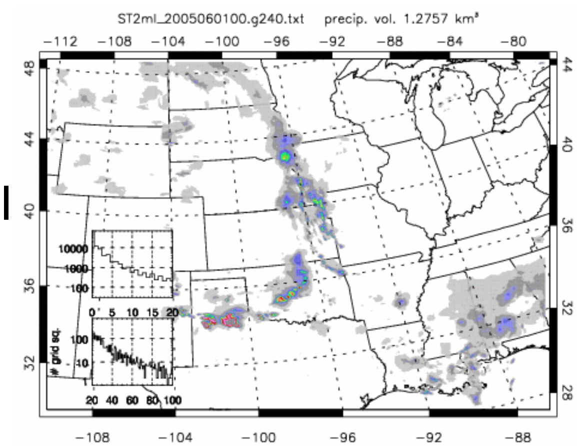
Spatial forecasts

Weather variables defined over spatial domains have **coherent spatial structure and features**

WRF model



Stage II radar



New spatial verification techniques aim to:

- account for field spatial structure
- provide information on error in physical terms
- account for uncertainties in location (and timing)

New spatial verification approaches

- Neighborhood (fuzzy) verification methods
  - give credit to "close" forecasts
- Scale decomposition methods
  - measure scale-dependent error
- Object-oriented methods
  - evaluate attributes of identifiable features
- Field verification
  - evaluate phase errors

Spatial Verification Intercomparison Project

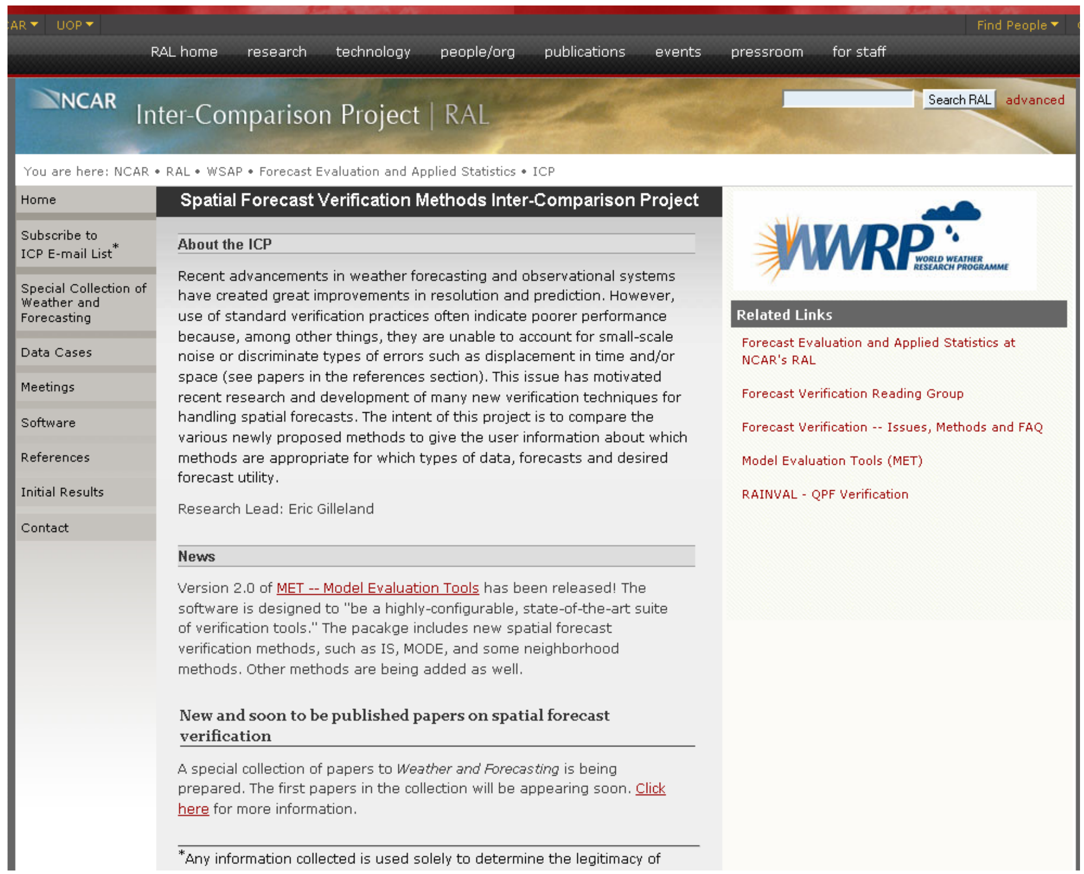
Begun February 2007

The main goals of this project are to:

- Obtain an inventory of the methods that are available and their capabilities
- Identify methods that
  - may be useful in operational settings
  - could provide automated feedback into forecasting systems
  - are particularly useful for specific applications (e.g., model diagnostics, hydrology, aviation)
- Identify where there may be holes in our capabilities and more research and development is needed

Spatial Verification Intercomparison Project

- http://www.ral.ucar.edu/projects/icp/index.html
  - Test cases
  - Results
  - Papers
  - Code



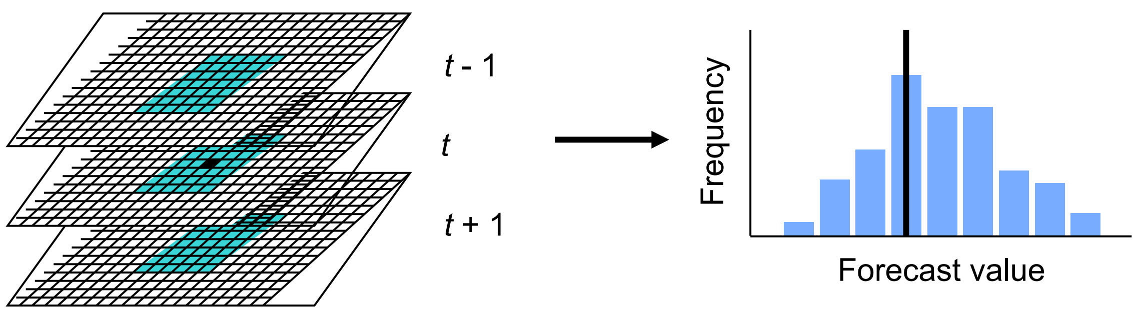
Neighborhood (fuzzy) verification methods

→ give credit to "close" forecasts

Neighborhood verification methods

- Don't require an exact match between forecasts and observations
  - Unpredictable scales
  - Uncertainty in observations

- Look in a space / time neighborhood around the point of interest



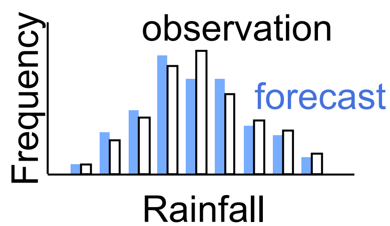
  - Evaluate using categorical, continuous, probabilistic scores / methods

Neighborhood verification methods

Treatment of forecast data within a window:

- Mean value (upscaling)
- Occurrence of event* somewhere in window
- Frequency of events in window → probability
- Distribution of values within window

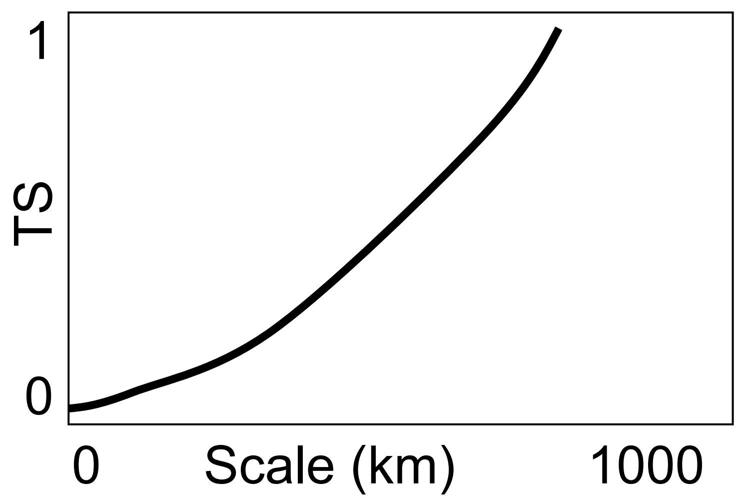
May also look in a neighborhood of observations



* *Event* defined as a value exceeding a given threshold, for example, rain exceeding 1 mm/hr

Oldest neighborhood verification method - upscaling

- Average the forecast and observations to successively larger grid resolutions, then verify using the usual metrics:
  - Continuous statistics – mean error, RMSE, correlation coefficient, etc.
  - Categorical statistics – POD, FAR, FBI, TS, ETS, etc.



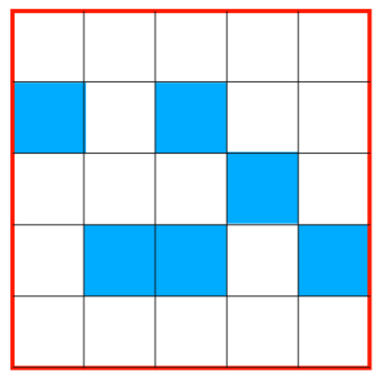
Fractions skill score

(Roberts and Lean, *MWR*, 2008)

- We want to know
  - How forecast skill varies with neighborhood size
  - The smallest neighborhood size that can be can be used to give sufficiently accurate forecasts
  - Does higher resolution NWP provide more accurate forecasts on scales of interest (e.g., river catchments)

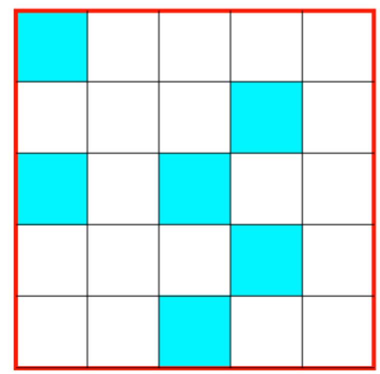
Compare forecast fractions with observed fractions (radar) in a *probabilistic* way over different sized neighbourhoods

$$FSS = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(P_{fcst} - P_{obs})^2}{\frac{1}{N}\sum_{i=1}^{N}P_{fcst}^{\ 2} + \frac{1}{N}\sum_{i=1}^{N}P_{obs}^{\ 2}}$$



Fraction = 6/25 = 0.24

observed

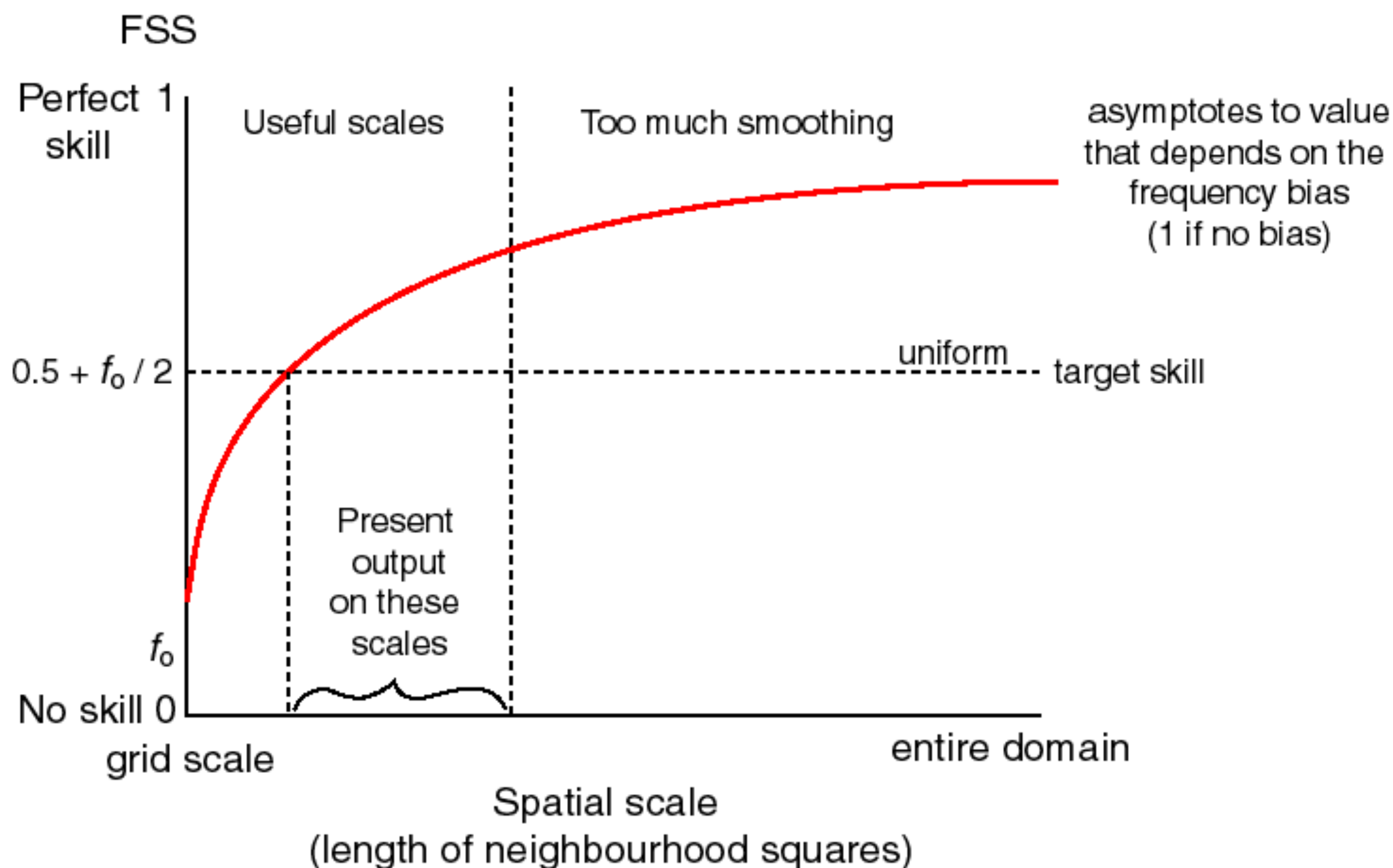


Fraction = 6/25 = 0.24

forecast

Fractions skill score

(Roberts and Lean, *MWR*, 2008)

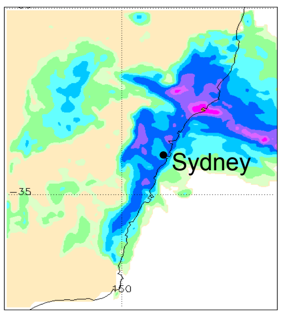


f_o=domain obs fraction

Spatial multi-event contingency table

Atger, *Proc. Nonlin. Geophys.*, 2001

- Experienced forecasters interpret output from a high resolution deterministic forecast in a *probabilistic* way



← "high probability of some heavy rain near Sydney",
not "62 mm of rain will fall in Sydney"

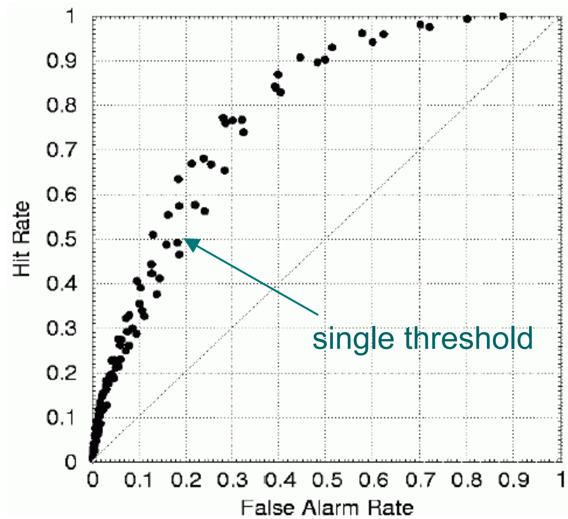
- The deterministic forecast is mentally "calibrated" according to how "close" the forecast is to the place / time / magnitude of interest.

Very close → high probability
Not very close → low probability

Spatial multi-event contingency table

Atger, *Proc. Nonlin. Geophys.*, 2001

- Verify using the Relative Operating Characteristic (ROC)



Measures how well the forecast can separate events from non-events based on some decision threshold

Decision thresholds to vary:

- magnitude (ex: 1 mm h^{-1} to 20 mm h^{-1})
- distance from point of interest (ex: within 10 km, ... , within 100 km)
- timing (ex: within 1 h, ... , within 12 h)
- anything else that may be important in interpreting the forecast

Different neighborhood verification methods have different decision models for what makes a *useful forecast*

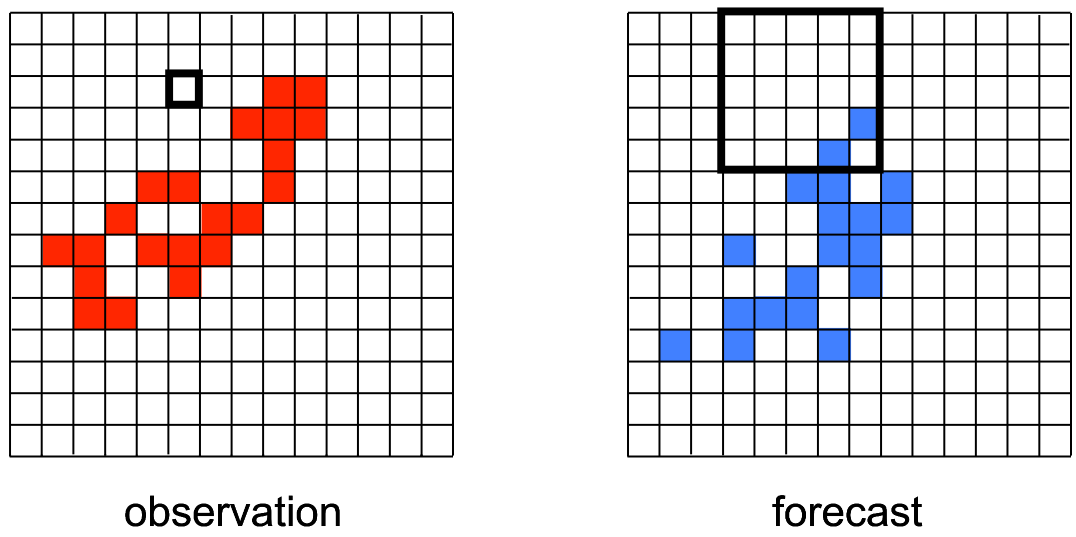
Neighborhood method	Matching strategy*	Decision model for useful forecast
Upscaling (Zepeda-Arce et al. 2000; Weygandt et al. 2004)	NO-NF	Resembles obs when averaged to coarser scales
Minimum coverage (Damrath 2004)	NO-NF	Predicts event over minimum fraction of region
Fuzzy logic (Damrath 2004), joint probability (Ebert 2002)	NO-NF	More correct than incorrect
Fractions skill score (Roberts and Lean 2008)	NO-NF	Similar frequency of forecast and observed events
Area-related RMSE (Rezacova et al. 2006)	NO-NF	Similar intensity distribution as observed
Pragmatic (Theis et al. 2005)	SO-NF	Can distinguish events and non-events
CSRR (Germann and Zawadzki 2004)	SO-NF	High probability of matching observed value
Multi-event contingency table (Atger 2001)	SO-NF	Predicts at least one event close to observed event
Practically perfect hindcast (Brooks et al. 1998)	SO-NF	Resembles forecast based on perfect knowledge of observations

*NO-NF = neighborhood observation-neighborhood forecast,
SO-NF = single observation-neighborhood forecast

from Ebert, *Meteorol. Appl.*, 2008

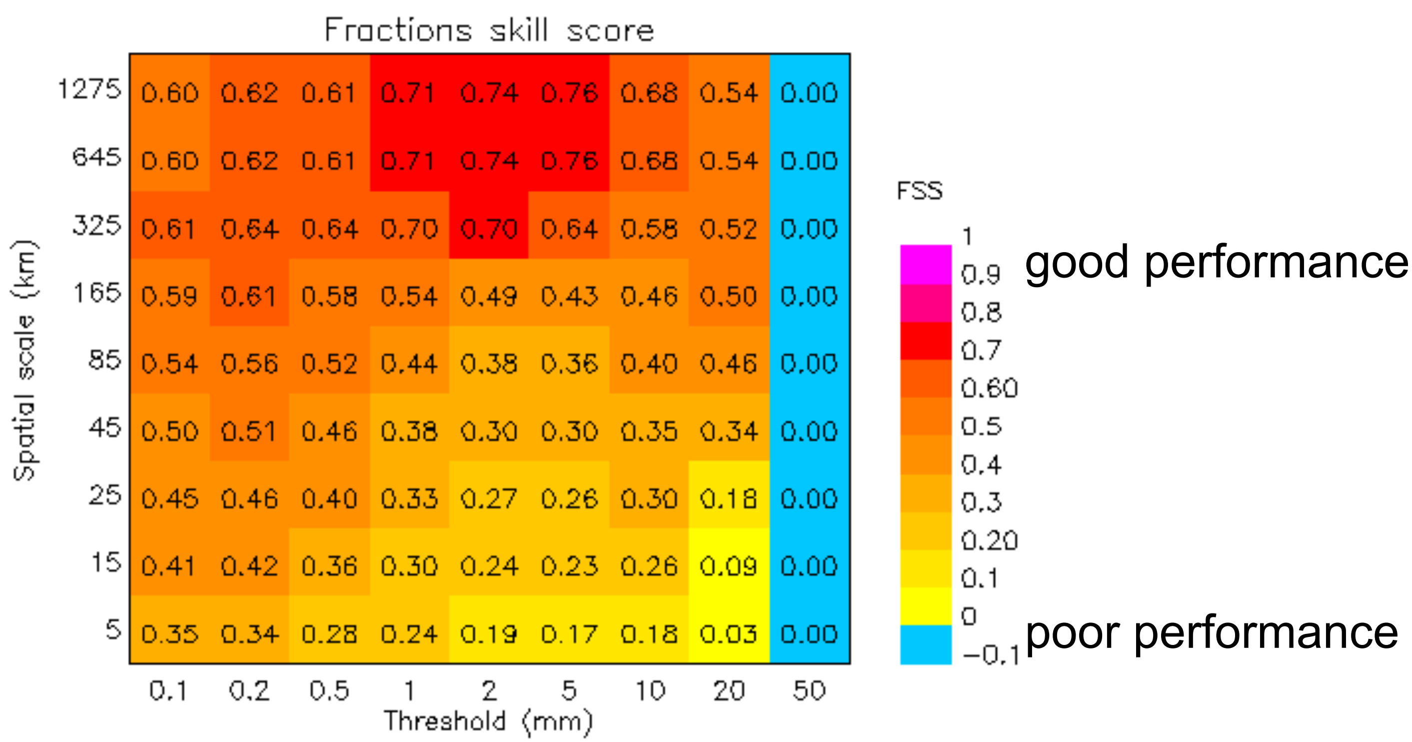
Moving windows

For each combination of neighborhood size and intensity threshold, accumulate scores as windows are moved through the domain

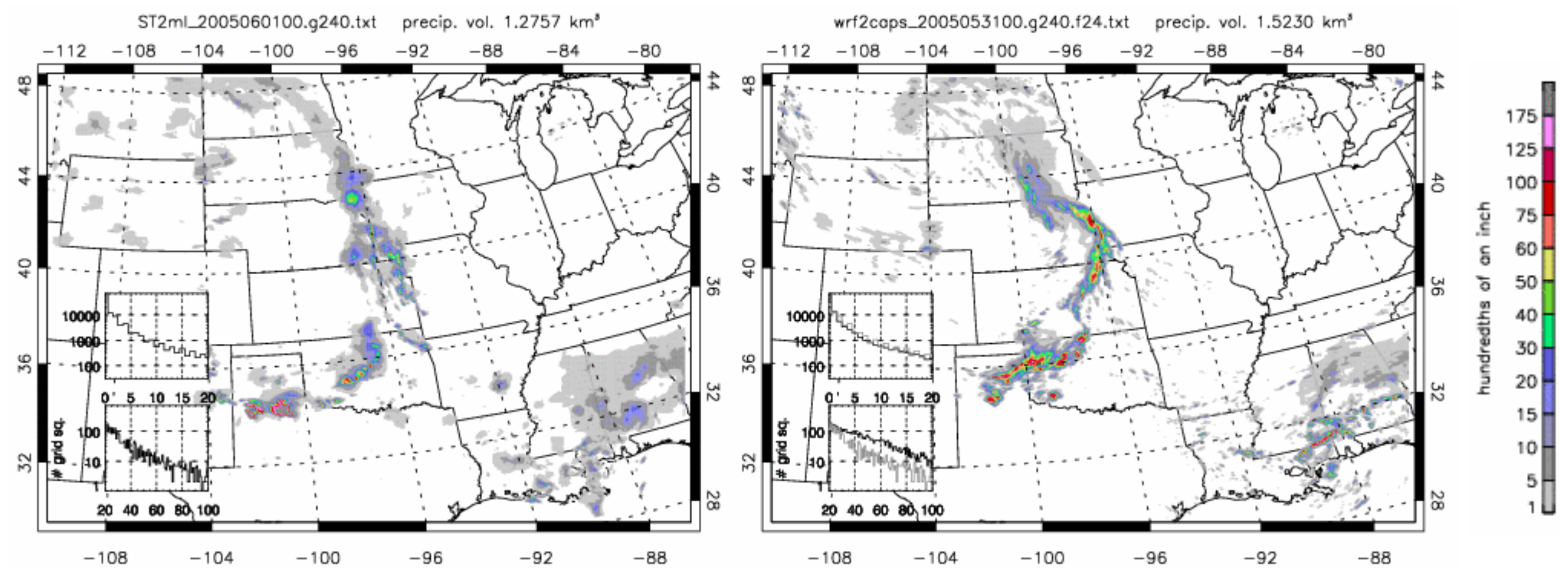


Multi-scale, multi-intensity approach

- Forecast performance depends on the scale and intensity of the event



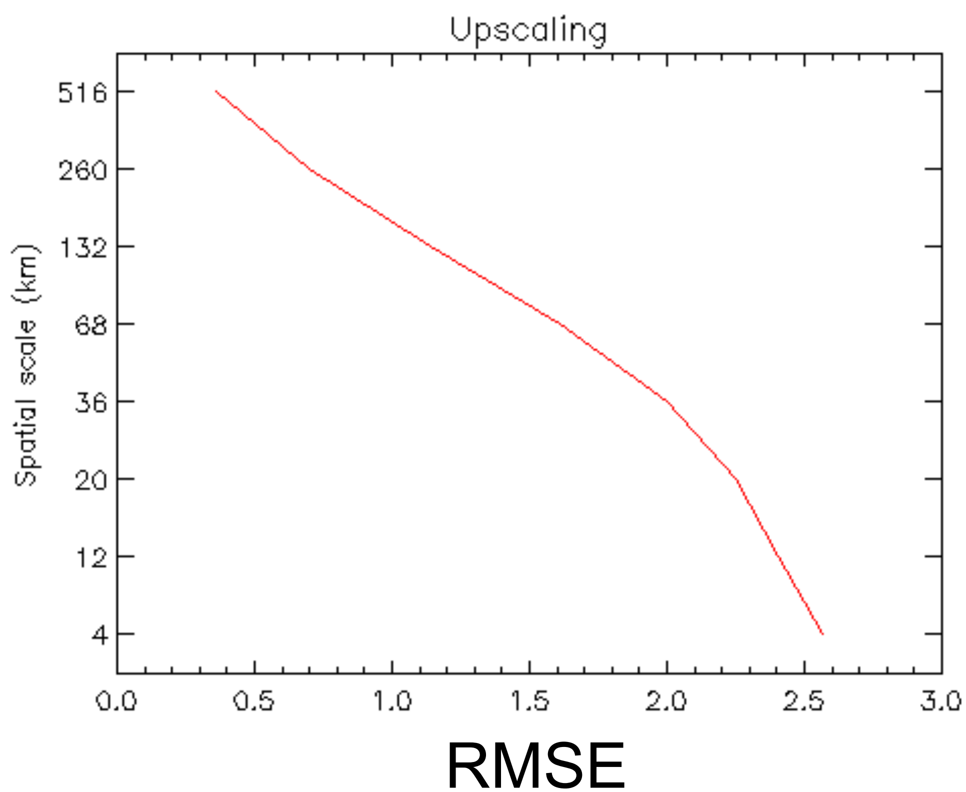
Example: Neighborhood verification of precipitation forecast over USA



1. How does the average forecast precipitation improve with increasing scale?
2. At which scales does the forecast rain distribution resemble the observed distribution?
3. How far away do we have to look to find at least one forecast value similar to the observed value?

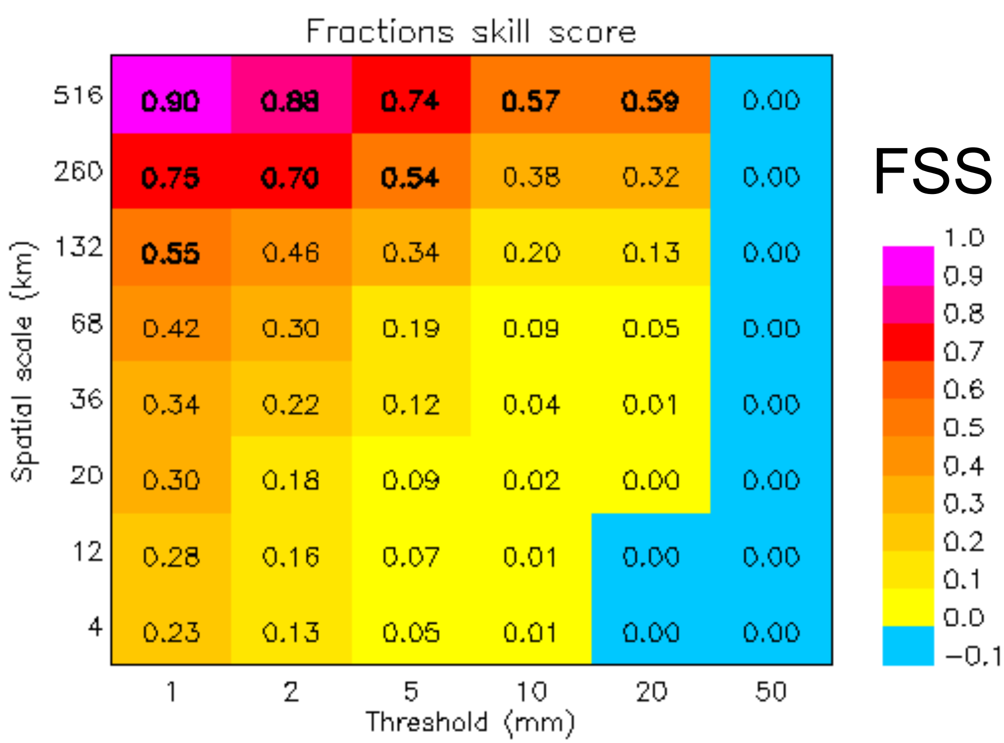
1. How does the average forecast precipitation improve with increasing scale?

- Upscaling method



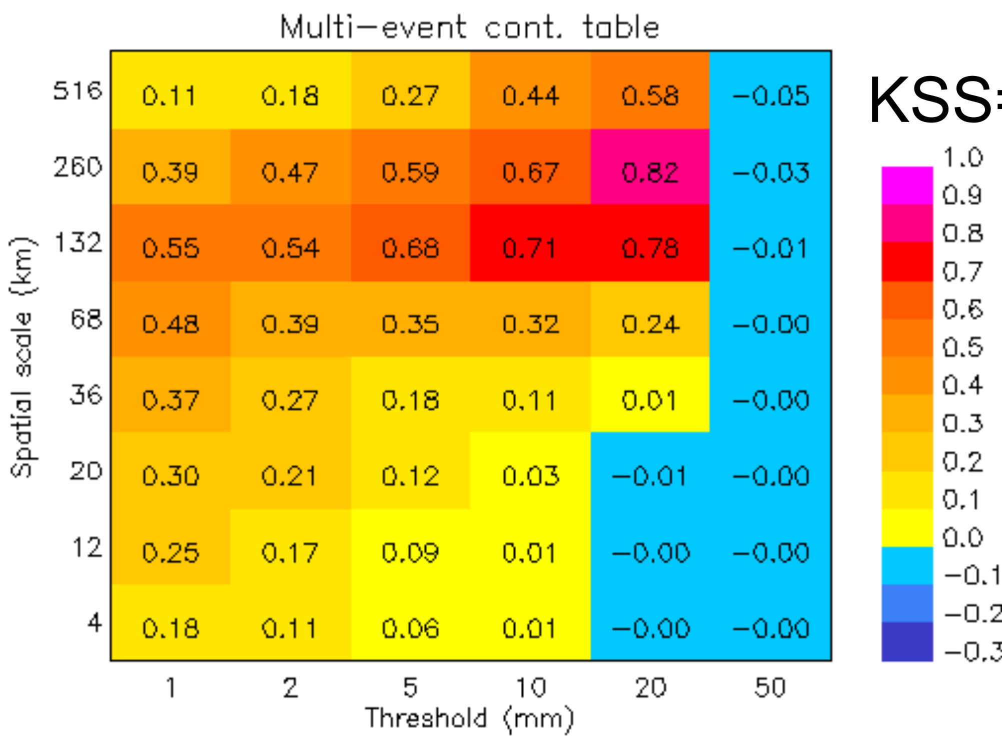
2. At which scales does the forecast rain distribution resemble the observed distribution?

- Fractions skill score



3. How far away do we have to look to find at least one forecast value similar to the observed value?

- Multi-event contingency table



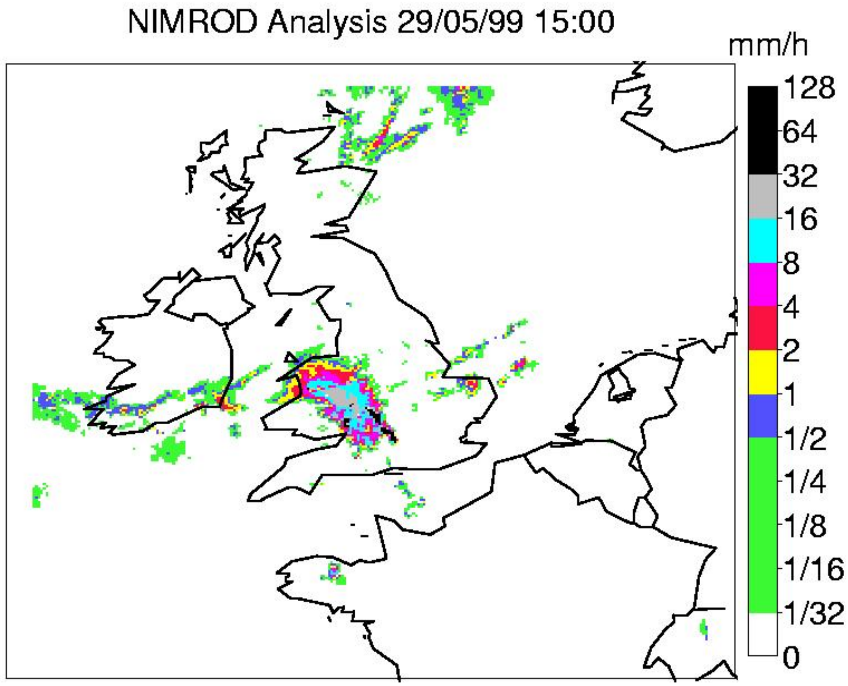
KSS=POD-POFD

Scale separation methods
→scale-dependent error

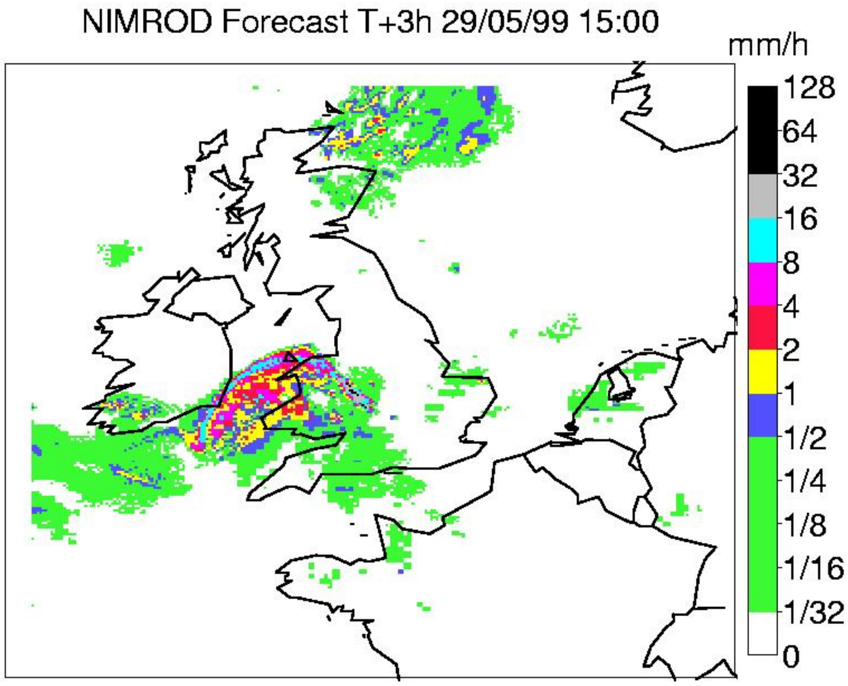
Intensity-scale method

Casati et al., *Met. Apps.*, 2004

Evaluate the forecast skill as a function of the **intensity** and the **spatial scale** of the error



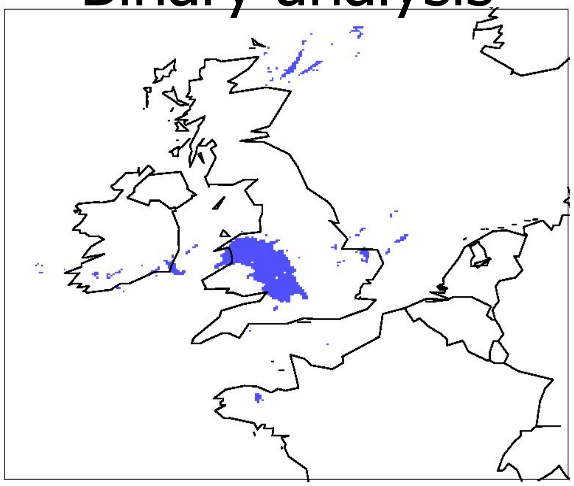
Precipitation analysis



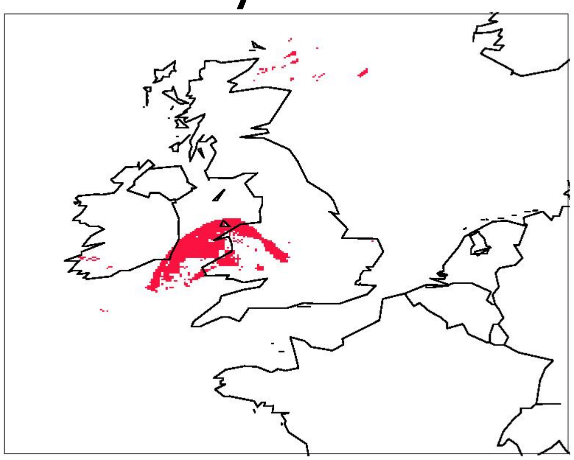
Precipitation forecast

Intensity threshold → binary images

Binary analysis

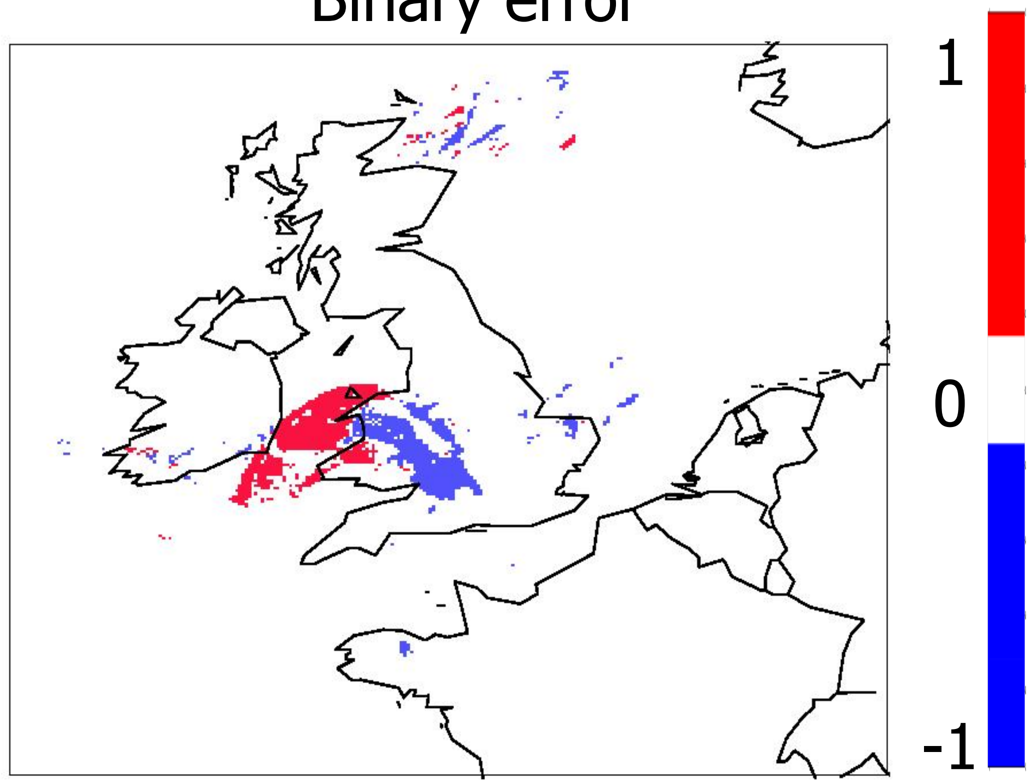


Binary forecast



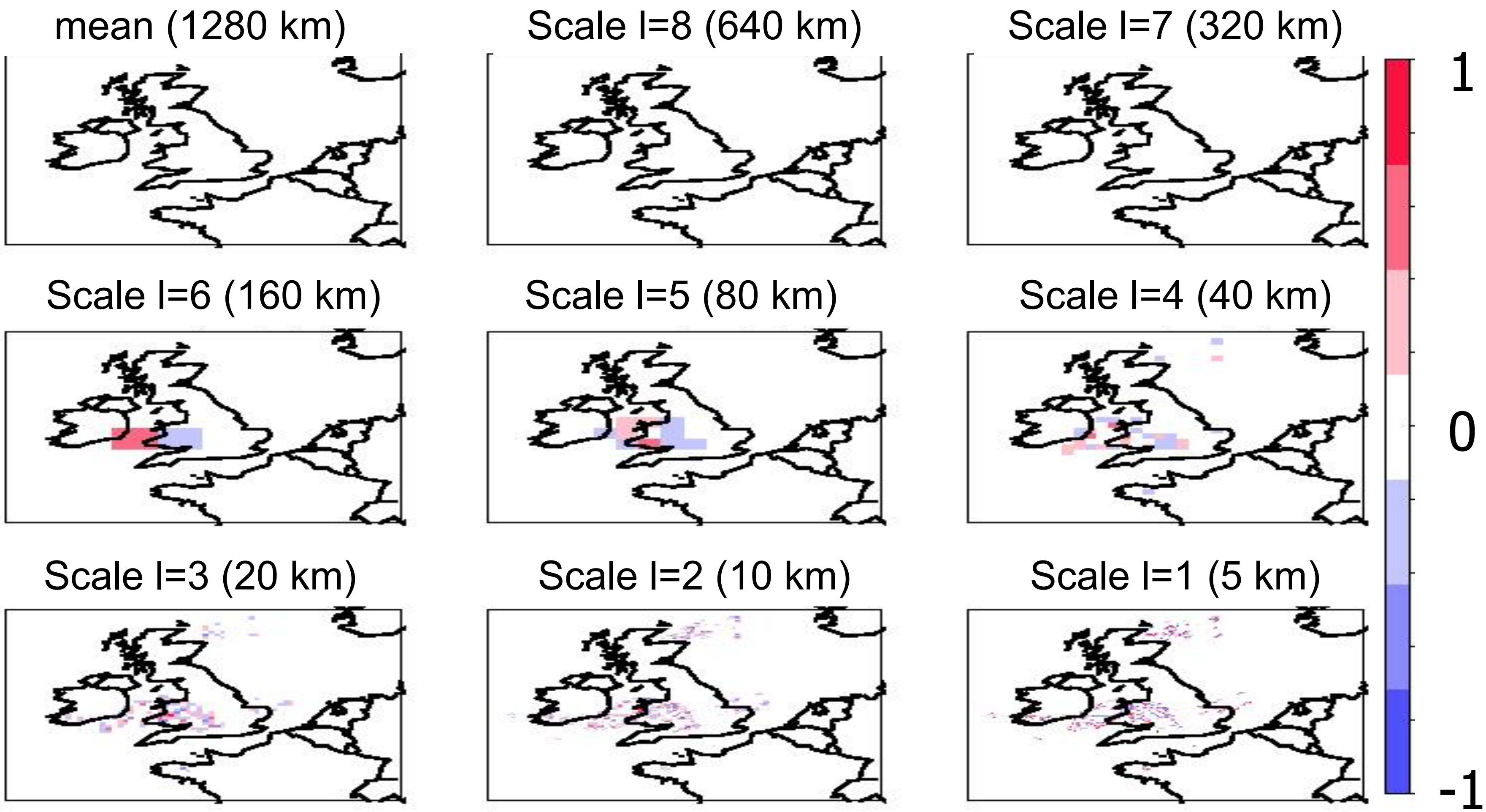
u=1 mm/h

Binary error



$$E_u = I_{Y'>u} - I_{X>u}$$

Scale → wavelet decomposition of binary error

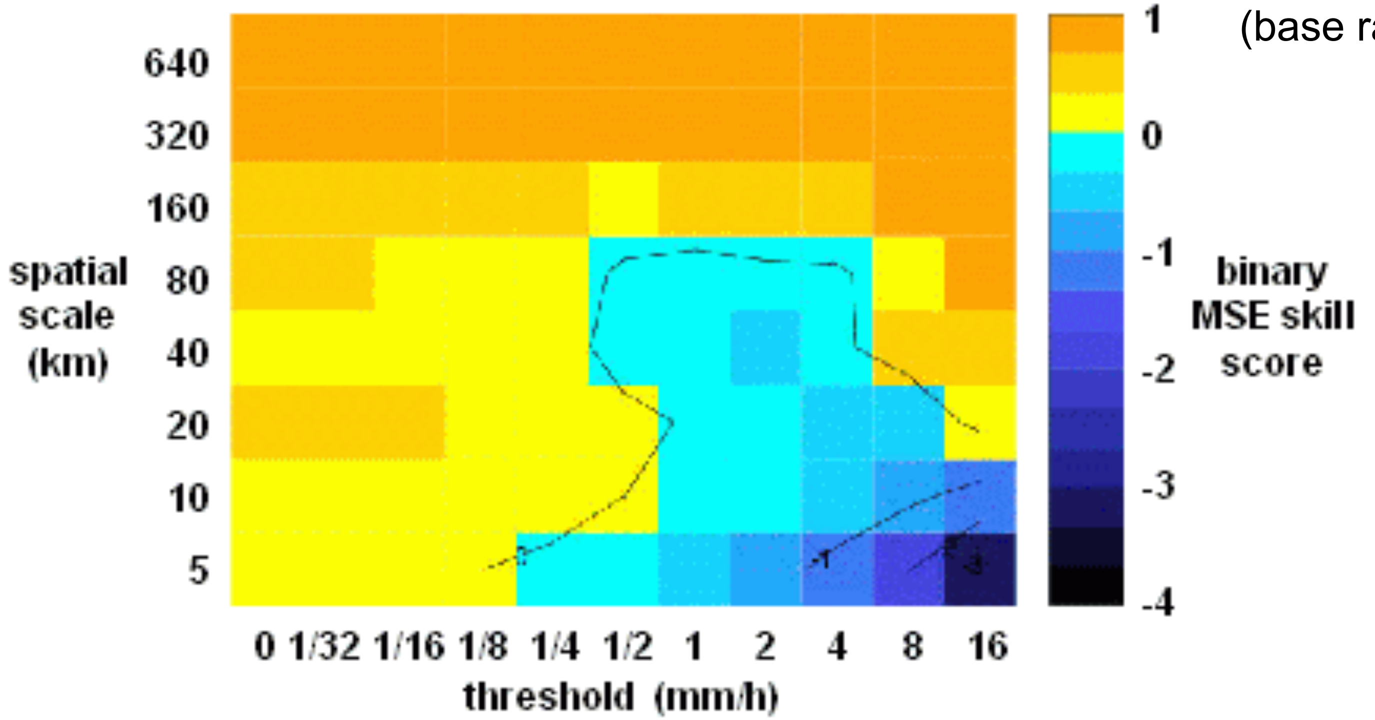


$$E_u = \sum_{l=1}^{L} E_{u,l} \qquad MSE_u = \sum_{l=1}^{L} MSE_{u,l}$$

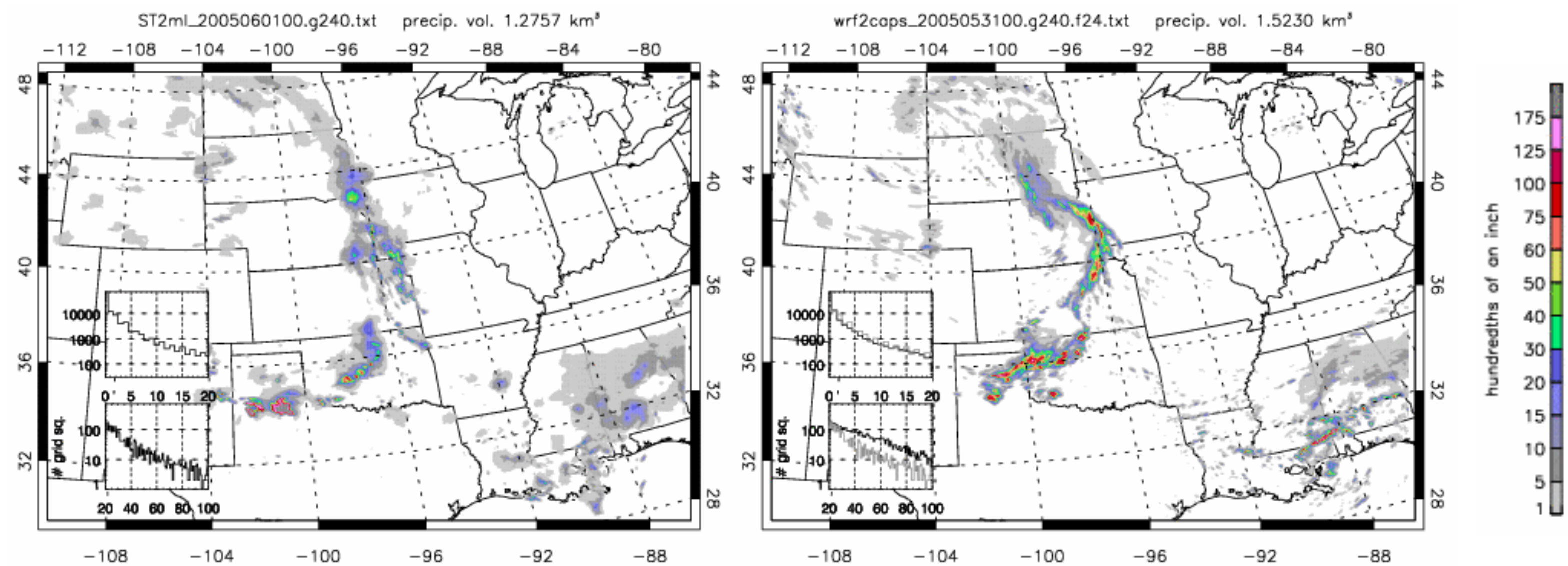
MSE skill score

$$SS_{u,l} = \frac{MSE_{u,l} - MSE_{u,l,random}}{MSE_{u,l,best} - MSE_{u,l,random}} = 1 - \frac{MSE_{u,l}}{2\varepsilon(1-\varepsilon)/L}$$

Sample climatology (base rate)

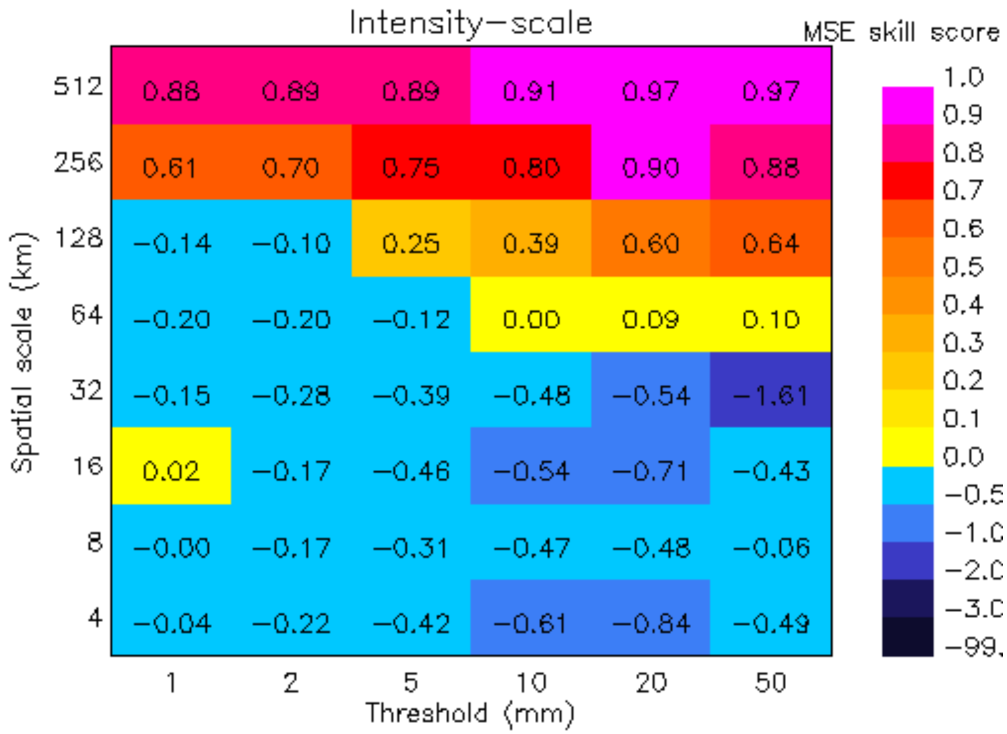


Example: Intensity-scale verification of precipitation forecast over USA



1. Which spatial scales are well represented and which scales have error?
2. How does the skill depend on the precipitation intensity?

Intensity-scale results



1. Which spatial scales are well represented and which scales have error?
2. How does the skill depend on the precipitation intensity?

What is the difference between neighborhood and scale decomposition approaches?

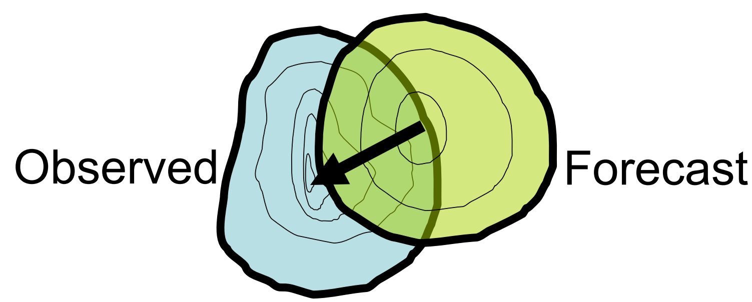
- Neighborhood (fuzzy) verification methods
  - Get scale information by ***filtering out higher resolution scales***
- Scale decomposition methods
  - Get scale information by ***isolating scales of interest***

Object-oriented methods
→evaluate attributes of features

Feature-based approach (CRA)

Ebert and McBride, *J. Hydrol.*, 2000

- Define entities using threshold (Contiguous Rain Areas)
- Horizontally translate the forecast until a *pattern matching* criterion is met:
  - minimum total squared error between forecast and observations
  - maximum correlation
  - maximum overlap
- The displacement is the vector difference between the original and final locations of the forecast.



CRA error decomposition

Total mean squared error (MSE)

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

The *displacement error* is the difference between the mean square error before and after translation

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

The *volume error* is the bias in mean intensity

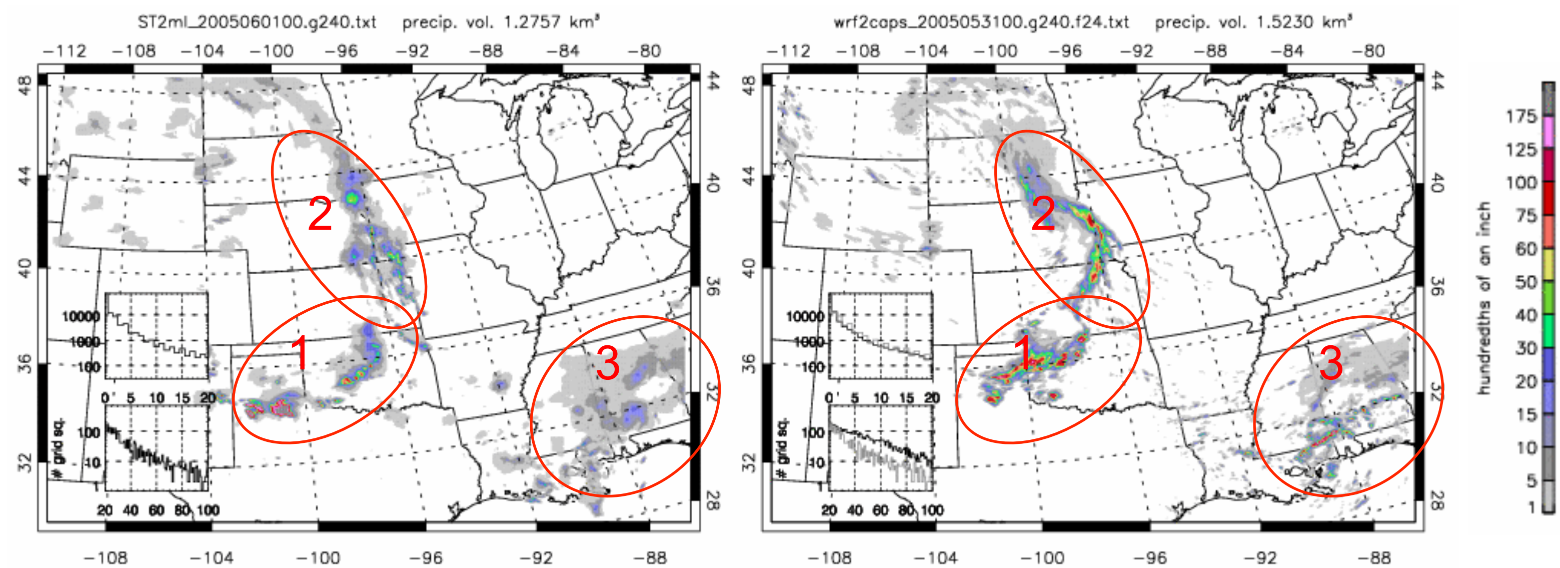
$$MSE_{volume} = (\overline{F} - \overline{X})^2$$

where $\overline{F}$ and $\overline{X}$ are the mean forecast and observed values after shifting.

The *pattern error*, computed as a residual, accounts for differences in the fine structure,

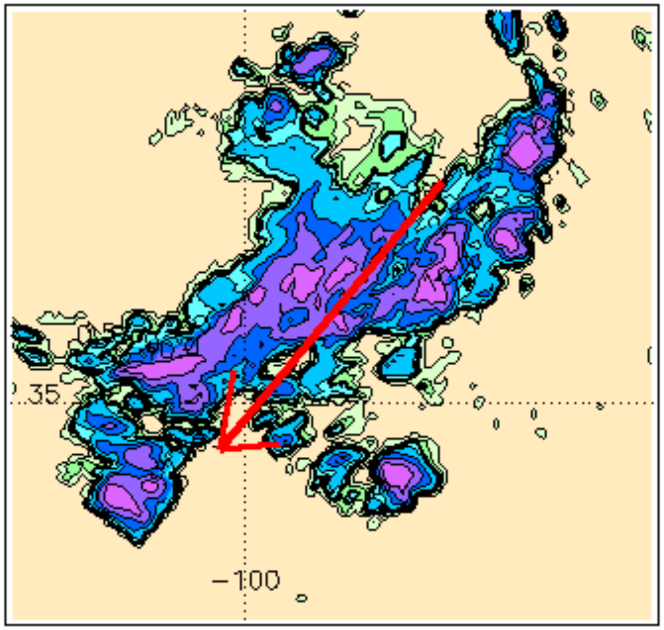
$$MSE_{pattern} = MSE_{shifted} - MSE_{volume}$$

Example: CRA verification of precipitation forecast over USA

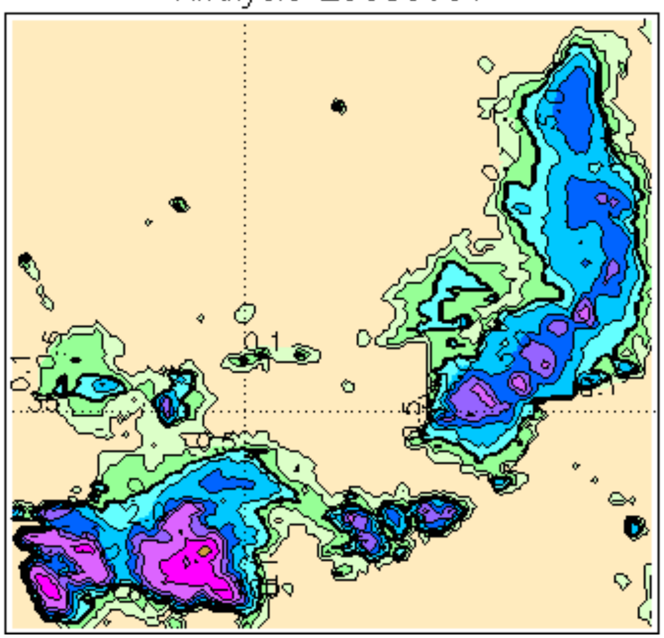


1. What is the location error of the forecast?
2. How do the forecast and observed rain areas compare? Average values? Maximum values?
3. How do the displacement, volume, and pattern errors contribute to the total error?

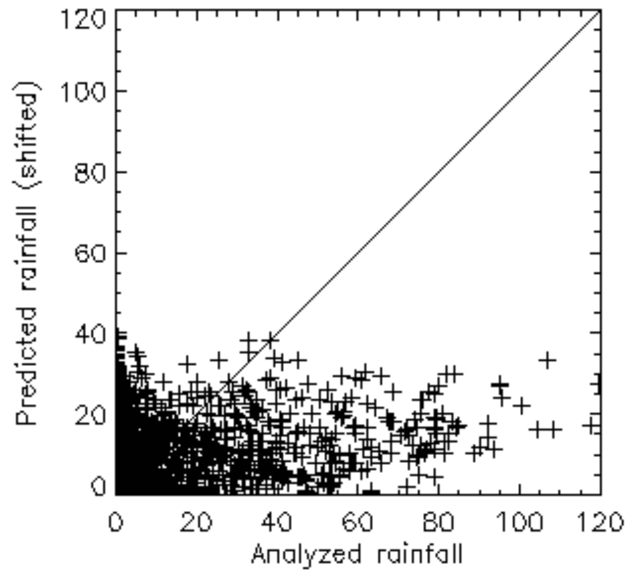
wrf2 fcst 20050601 hour 00–24



Analysis 20050601



CRA 20050601



wrf2 24h fcst 20050601 n=8423
(33.49°,–102.28°) to (37.77°,–96.00°)
Verif. grid=0.042° CRA threshold=1.0 mm/h

	Analysed	Forecast
# gridpoints ≥1 mm/h	3304	3597
Average rainrate (mm/h)	3.58	3.61
Maximum rain (mm/h)	119.63	39.12
Rain volume (km^3)	0.51	0.52

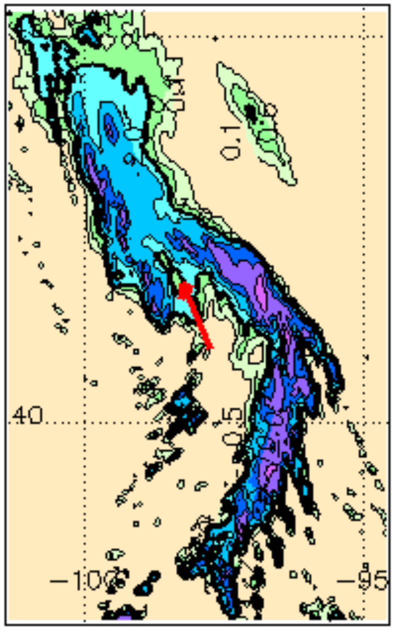
Displacement (E,N) = [2.20°,1.92°] max.corr matching

	Original	Shifted
RMS error (mm/d)	12.81	10.24
Correlation coefficient	–0.167	0.305

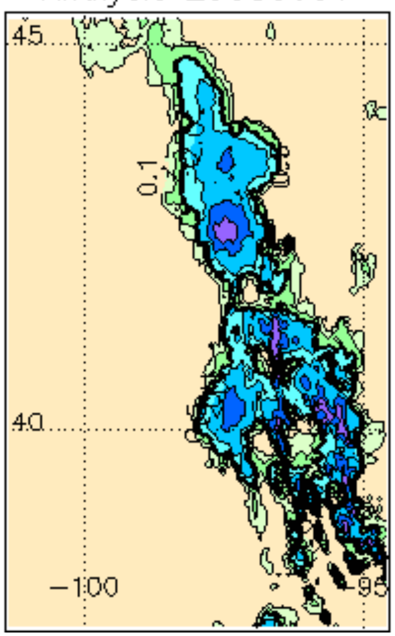
Error Decomposition:

Displacement error	36.1%
Volume error	0.0%
Pattern error	63.9%

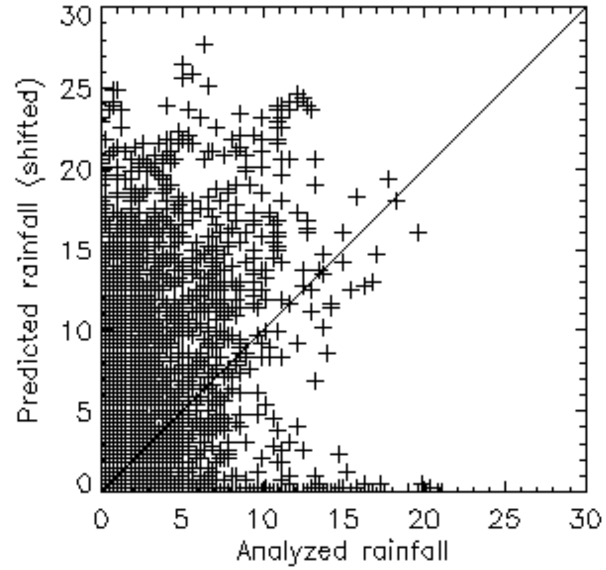
wrf2 fcst 20050601 hour 00–24



Analysis 20050601



CRA 20050601



wrf2 24h fcst 20050601 n=11007
(37.52°,−101.29°) to (45.29°,−94.65°)
Verif. grid=0.042° CRA threshold=1.0 mm/h

	Analysed	Forecast
# gridpoints ≥1 mm/h	4840	5699
Average rainrate (mm/h)	1.52	2.68
Maximum rain (mm/h)	21.08	27.69
Rain volume (km^3)	0.26	0.46

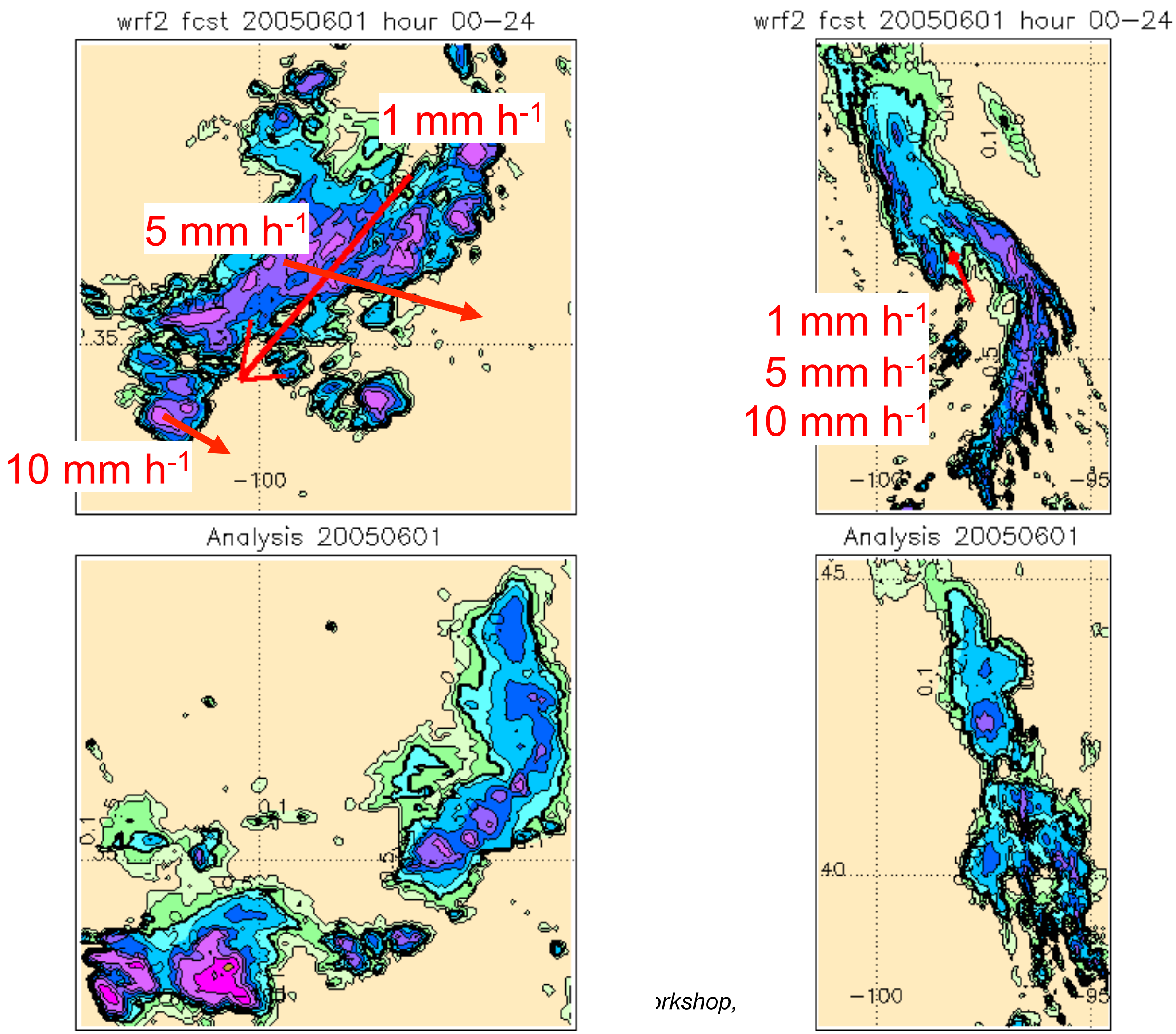
Displacement (E,N) = [0.52°,−0.84°] max.corr matching

	Original	Shifted
RMS error (mm/d)	5.11	4.65
Correlation coefficient	−0.040	0.193

Error Decomposition:

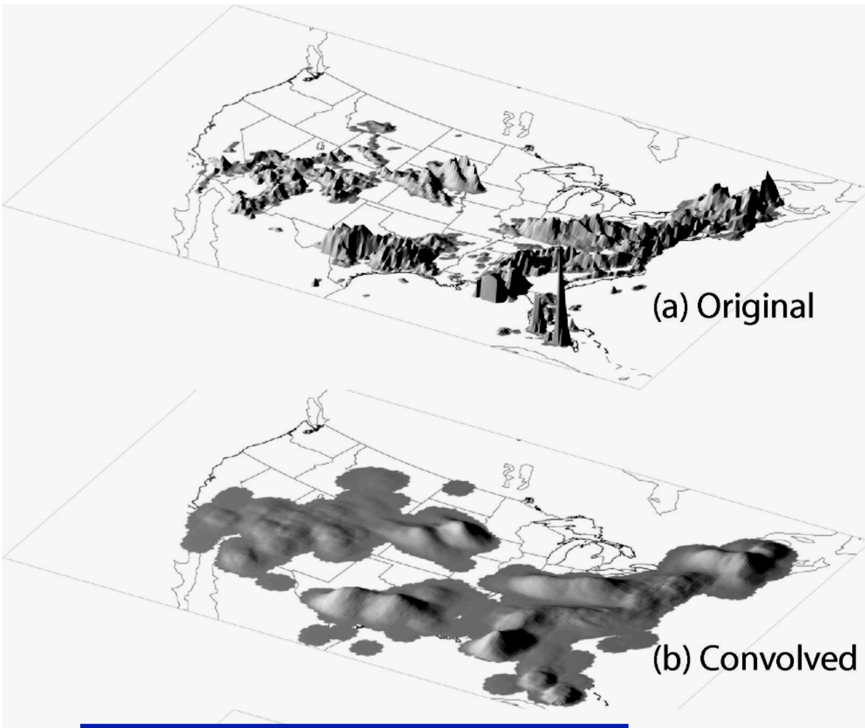
Displacement error	18.7%
Volume error	4.9%
Pattern error	76.4%

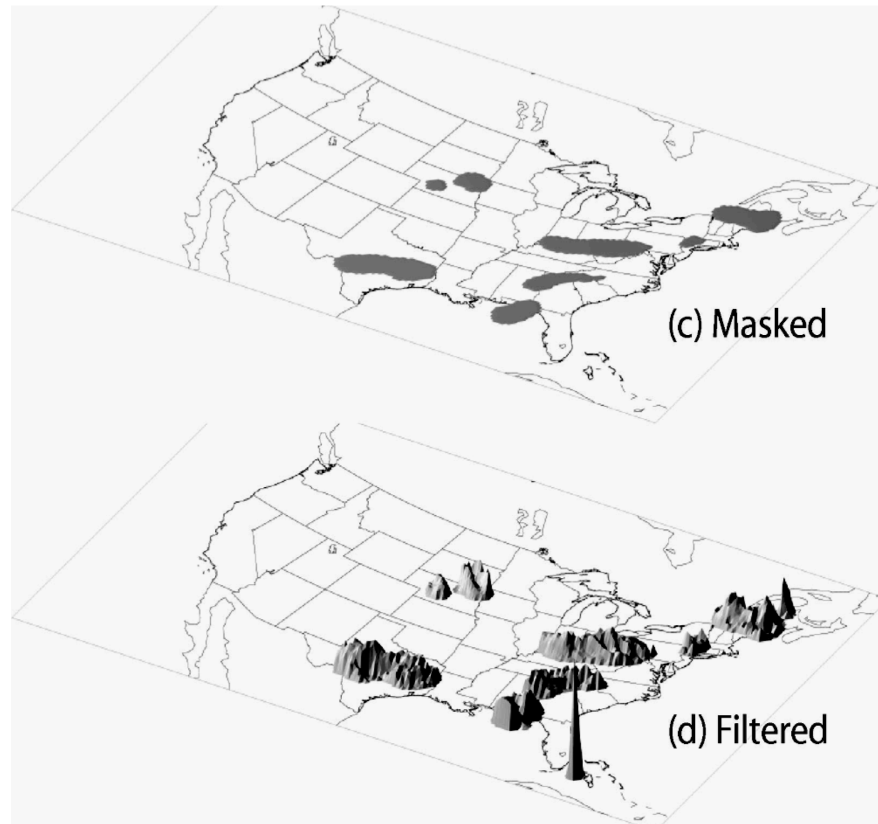
Sensitivity to rain threshold



MODE – Method for Object-based Diagnostic Evaluation

Davis et al., *MWR*, 2006

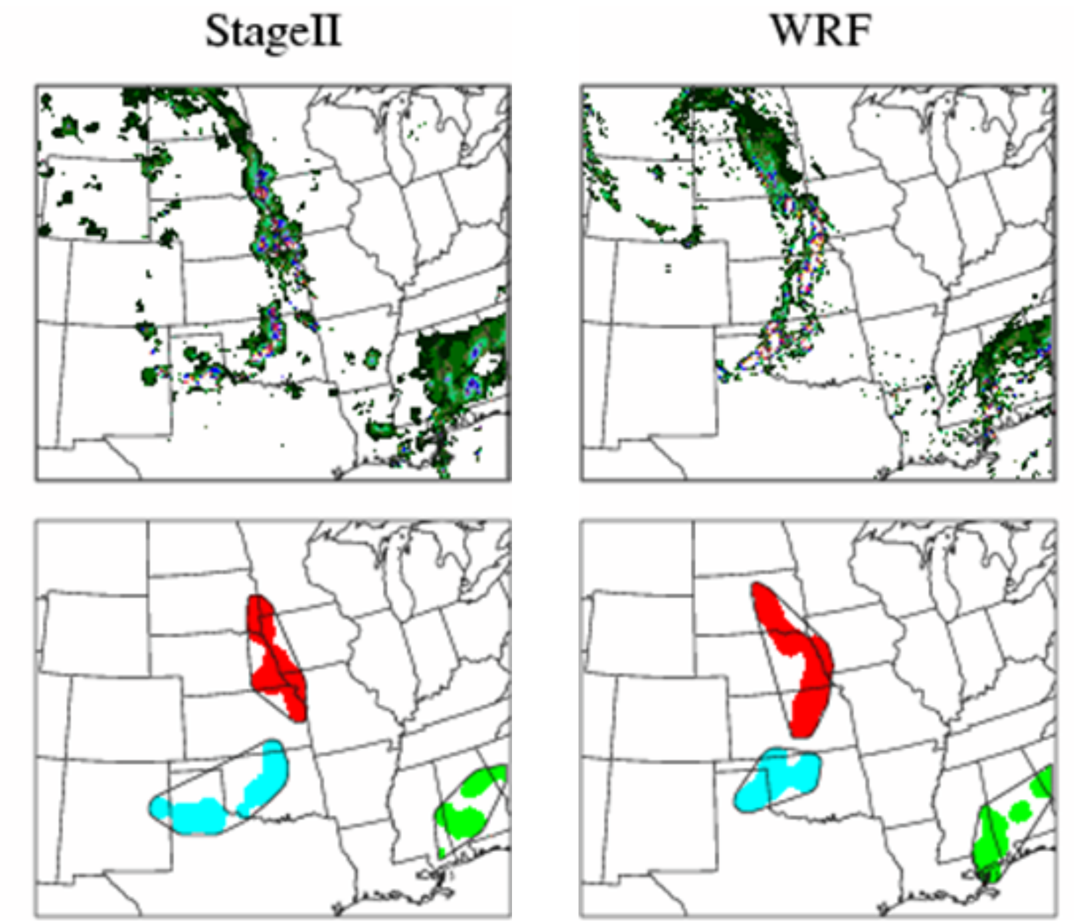




Two parameters:

1. Convolution radius
2. Threshold

MODE object matching/merging



24h forecast of 1h rainfall on 1 June 2005

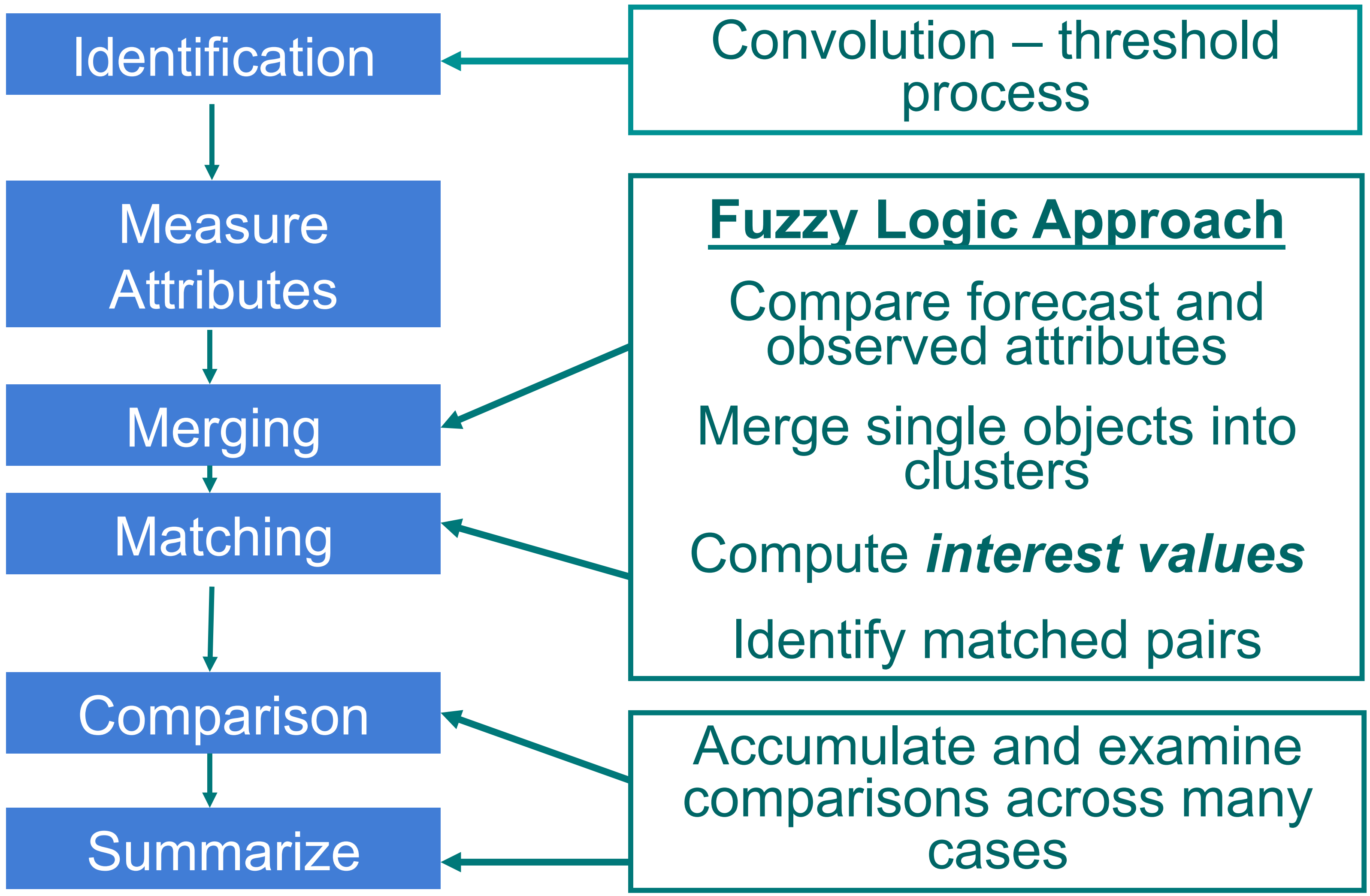
Compare attributes:

- centroid location
- intensity distribution
- area
- orientation
- etc.

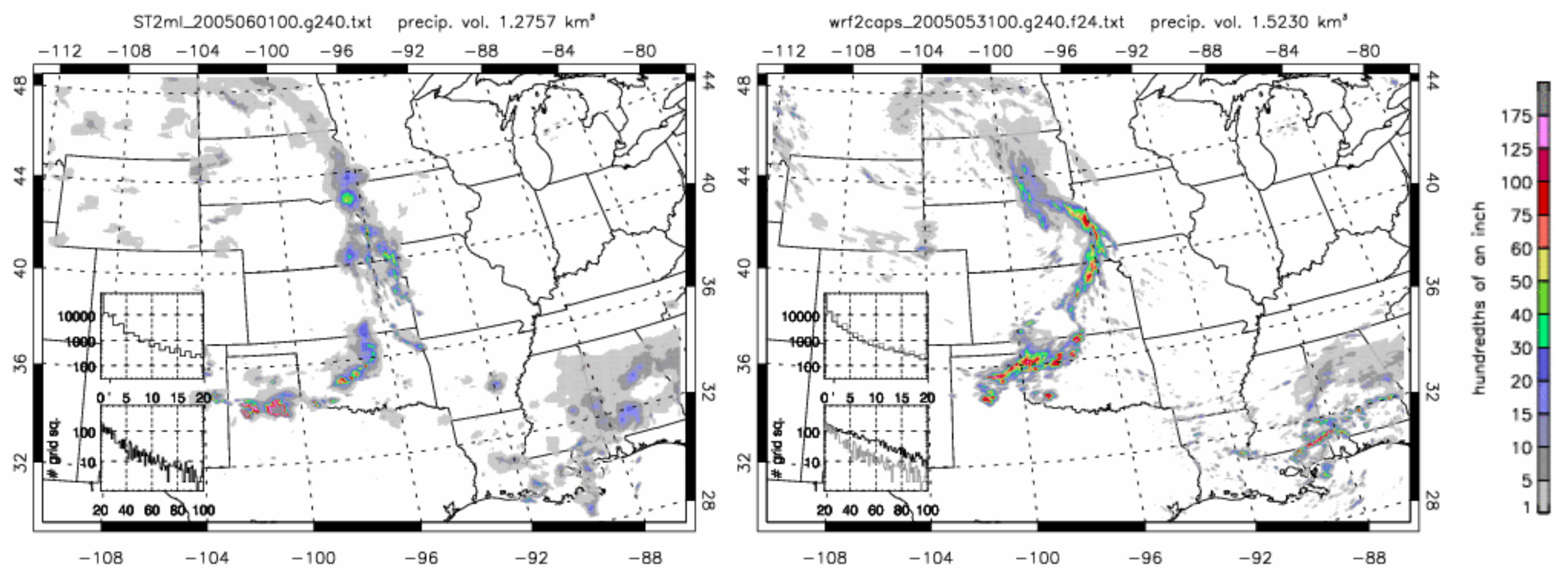
When objects not matched:

- false alarms
- missed events
- rain volume
- etc.

MODE methodology



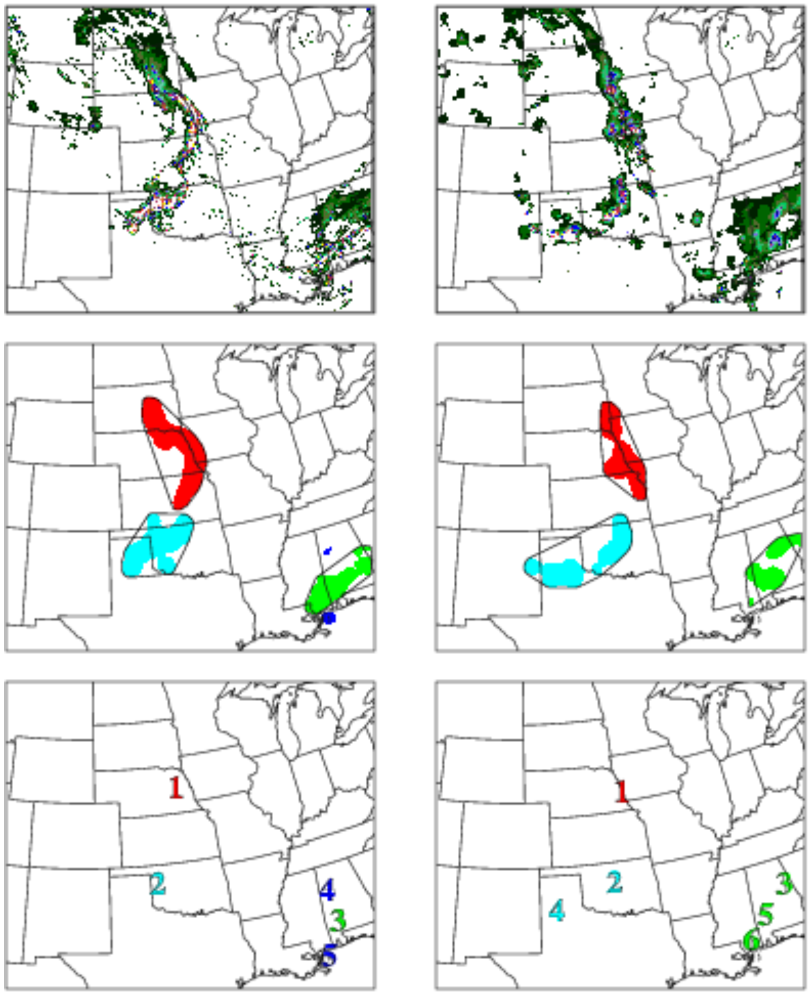
Example: MODE verification of precipitation forecast over USA



1. What is the location error of the forecast?
2. How do the forecast and observed rain areas compare? Average values? Maximum values? Shape?
3. What is the overall quality of the forecast as measured by the median of the maximum object interest values?

MODE applied to our US rain example

WRF StageII



WRF	StageII	Interest
1	1	0.9665
3	5	0.9262
2	2	0.9097
3	6	0.8715
2	4	0.8494
3	3	0.6808
4	3	0.6187
5	5	0.6138
1	2	0.6030
5	6	0.5992
2	1	0.5991
4	5	0.5886
4	6	0.5484
5	3	0.4399
4	1	N/A
1	4	N/A
3	2	N/A
3	4	N/A
4	4	N/A
5	4	N/A
1	5	N/A
2	5	N/A
4	2	N/A
5	2	N/A
1	3	N/A
1	6	N/A
2	6	N/A
2	3	N/A
5	1	N/A
3	1	N/A

Displacement errors

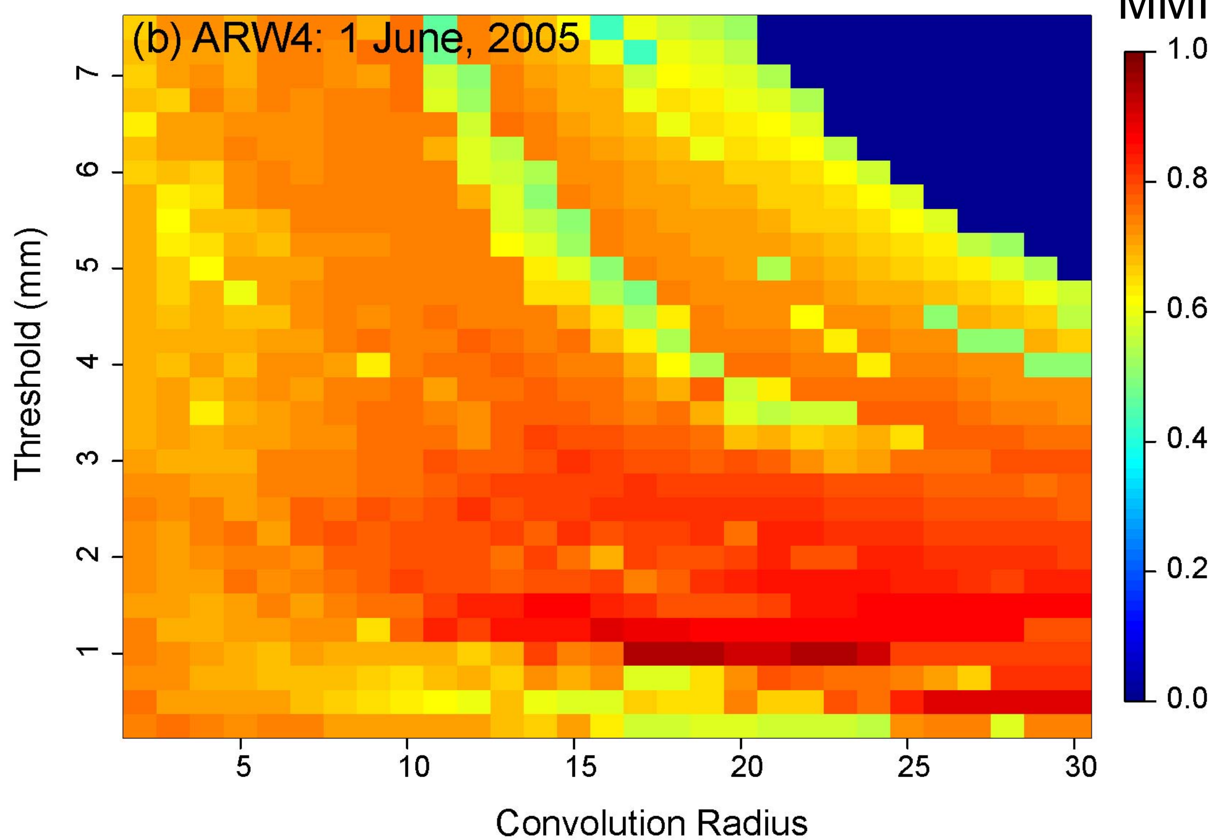
1 25 km

2 23 km

3 30 km

Issue Time:	May 31, 2005 00:00:00		WRF	StageII
Valid Time:	Jun 1, 2005 00:00:00	Raw Thresh:	0.00 in/100	0.00 in/100
Lead Time:	24 hours	Mask Bad:	off	off
Accum Time:	1 hours	Conv Radius:	15 gs	15 gs
Fuzzy Engine Weights		Conv Thresh:	5.00 in/100	5.00 in/100

Sensitivity to rain threshold and convolution radius



MMI = median of maximum interest (overall goodness of fit)

(Note: This is not for the same case)

Structure-Amplitude-Location (SAL)

Wernli et al., *Mon. Wea. Rev.*, 2008

For a chosen domain and precipitation threshold, compute:

Amplitude error $A = (D(R_{fcst}) - D(R_{obs})) / 0.5*(D(R_{fcst}) + D(R_{obs}))$

D(…) denotes the area-mean value (e.g., catchment)
$A \in [-2, \ldots, 0, \ldots, +2]$

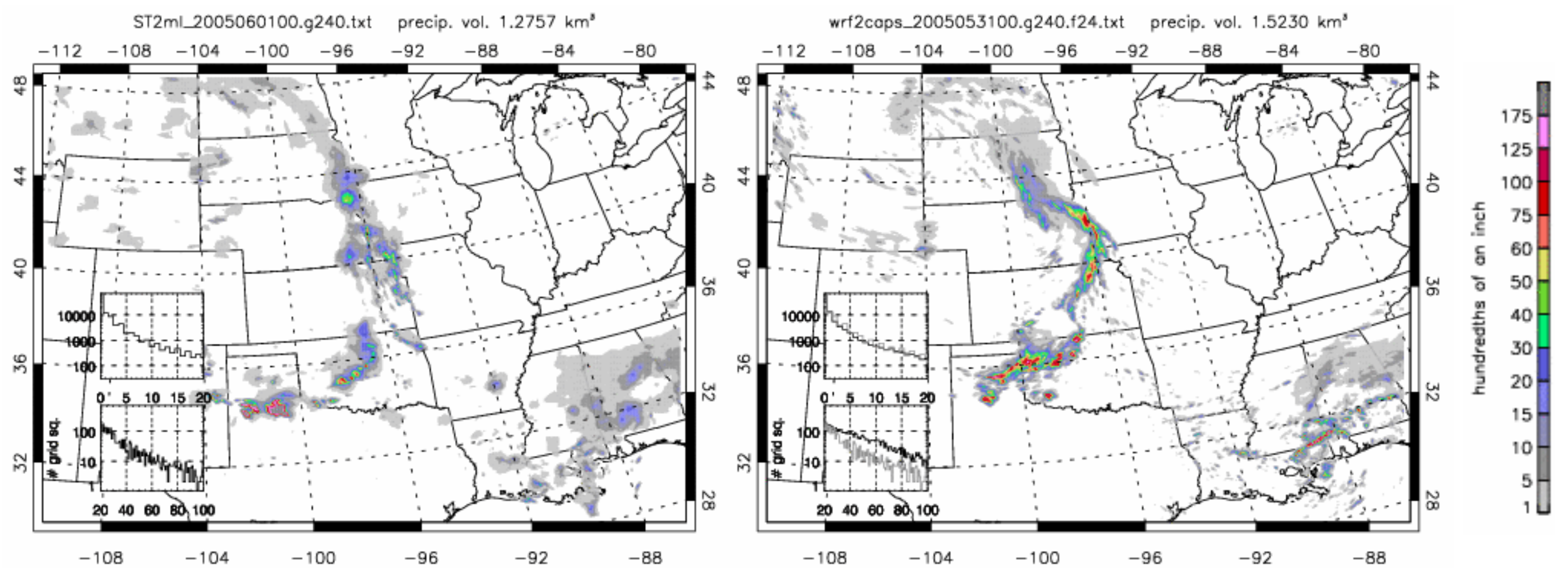
Location error $L = |r(R_{fcst}) - r(R_{obs})| / dist_{max}$

r(…) denotes the centre of mass of the precipitation field in the area
$L \in [0, \ldots, 1]$

Structure error $S = (V(R_{fcst}^*) - V(R_{obs}^*)) / 0.5*(V(R_{fcst}^*) + V(R_{obs}^*))$

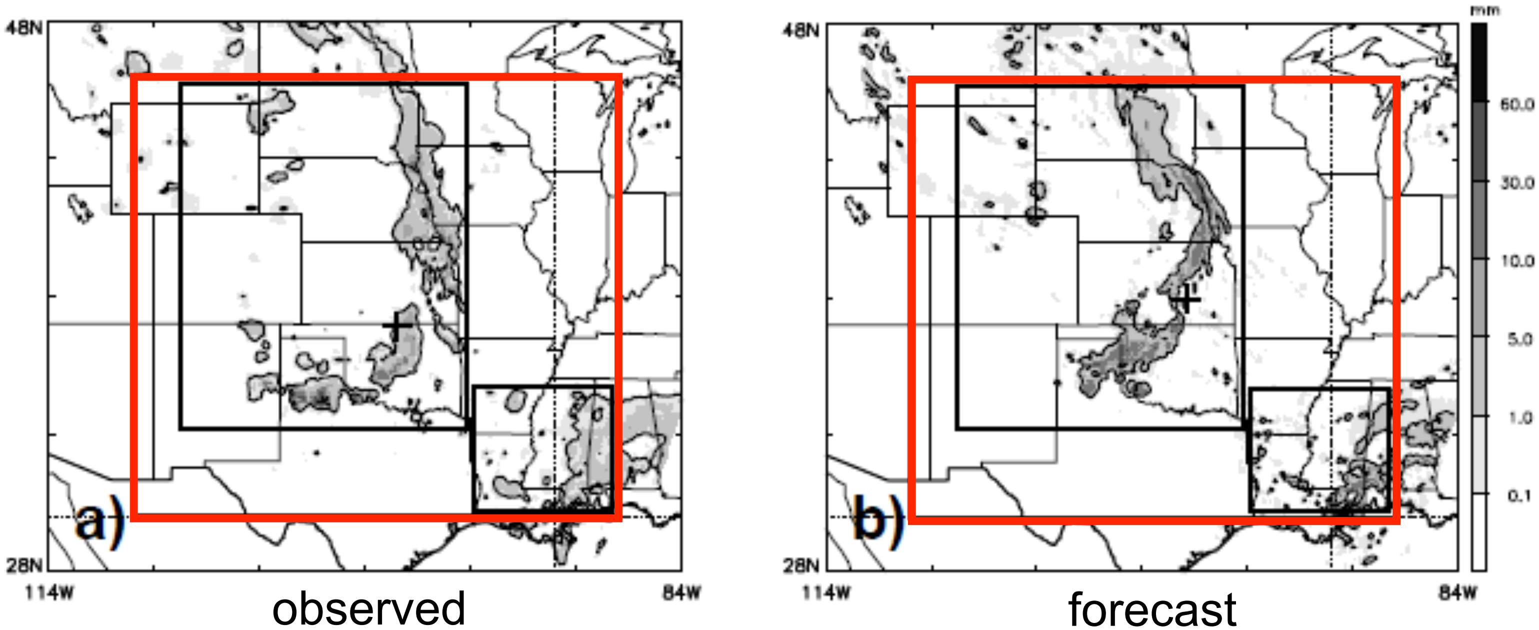
V(…) denotes the weighted volume average of all scaled precipitation objects in considered area, $R^* = R / R_{max}$
$S \in [-2, \ldots, 0, \ldots, +2]$

Example: SAL verification of precipitation forecast over USA



1. Is the domain average precipitation correctly forecast?
2. Is the mean location of the precipitation distribution in the domain correctly forecast?
3. Does the forecast capture the typical structure of the precipitation field (e.g., large broad objects vs. small peaked objects)?

SAL verification results



1. Is the domain average precipitation correctly forecast? A = 0.21
2. Is the mean location of the precipitation distribution in the domain correctly forecast? L = 0.06
3. Does the forecast capture the typical structure of the precipitation field (e.g., large broad objects vs. small peaked objects)? S = 0.46

(perfect=0)

Field verification

→ evaluate phase errors

Displacement and Amplitude Score (DAS)

Keil and Craig, *WAF*, 2009

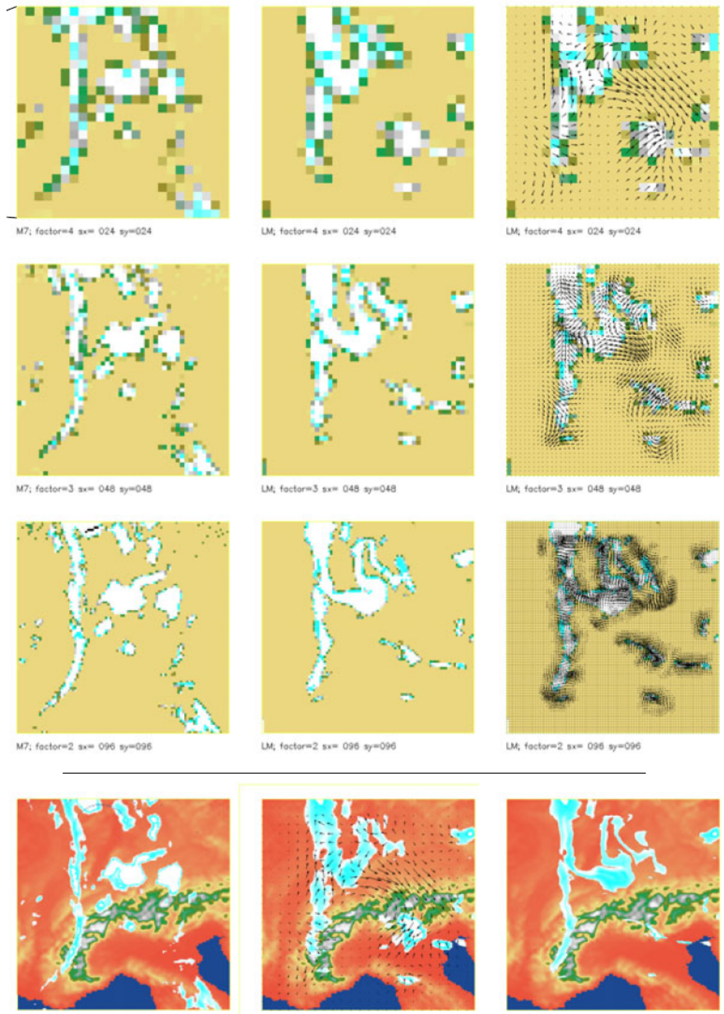
Combines distance and amplitude measures by matching forecast → observation & observation → forecast

- Pyramidal image matching (optical flow) to get vector displacement field → *DIS*
- Intensity errors for morphed field → *AMP*
- Displacement-amplitude score

$$DAS = \frac{DIS}{D_{max}} + \frac{AMP}{I_0}$$

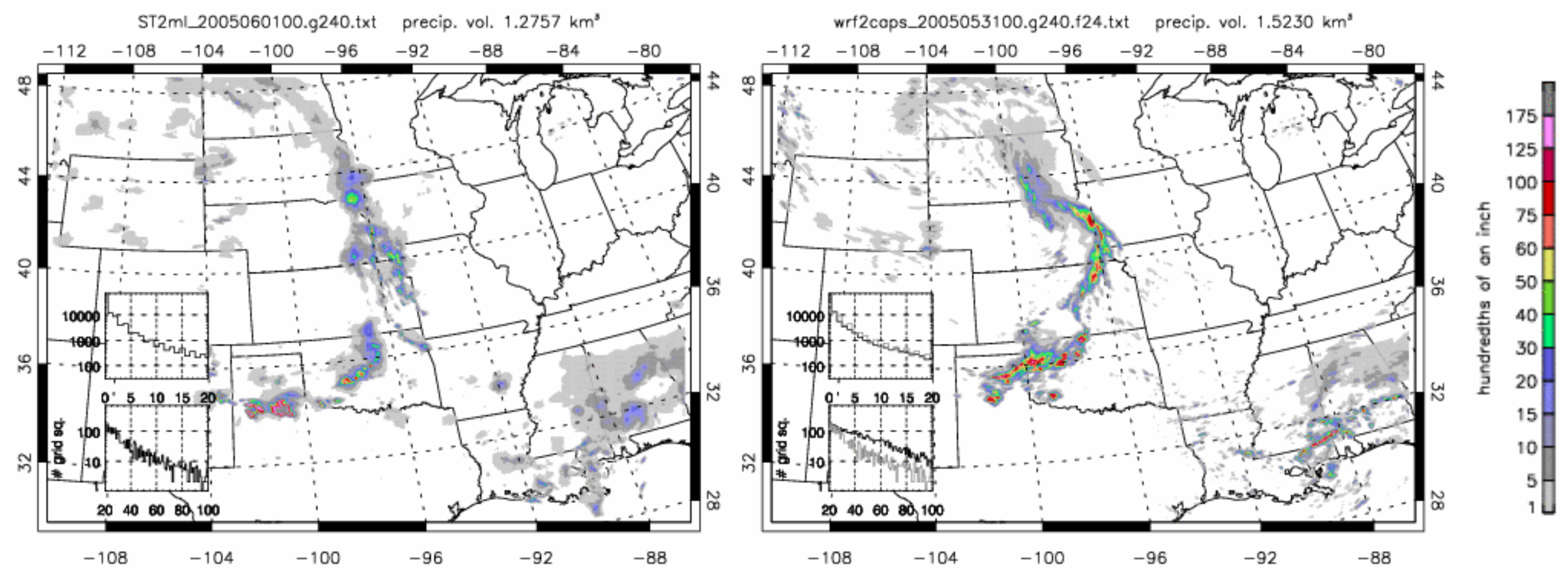
Morphing example (old)

Meteosat 7 | LM | LM + displ. vectors



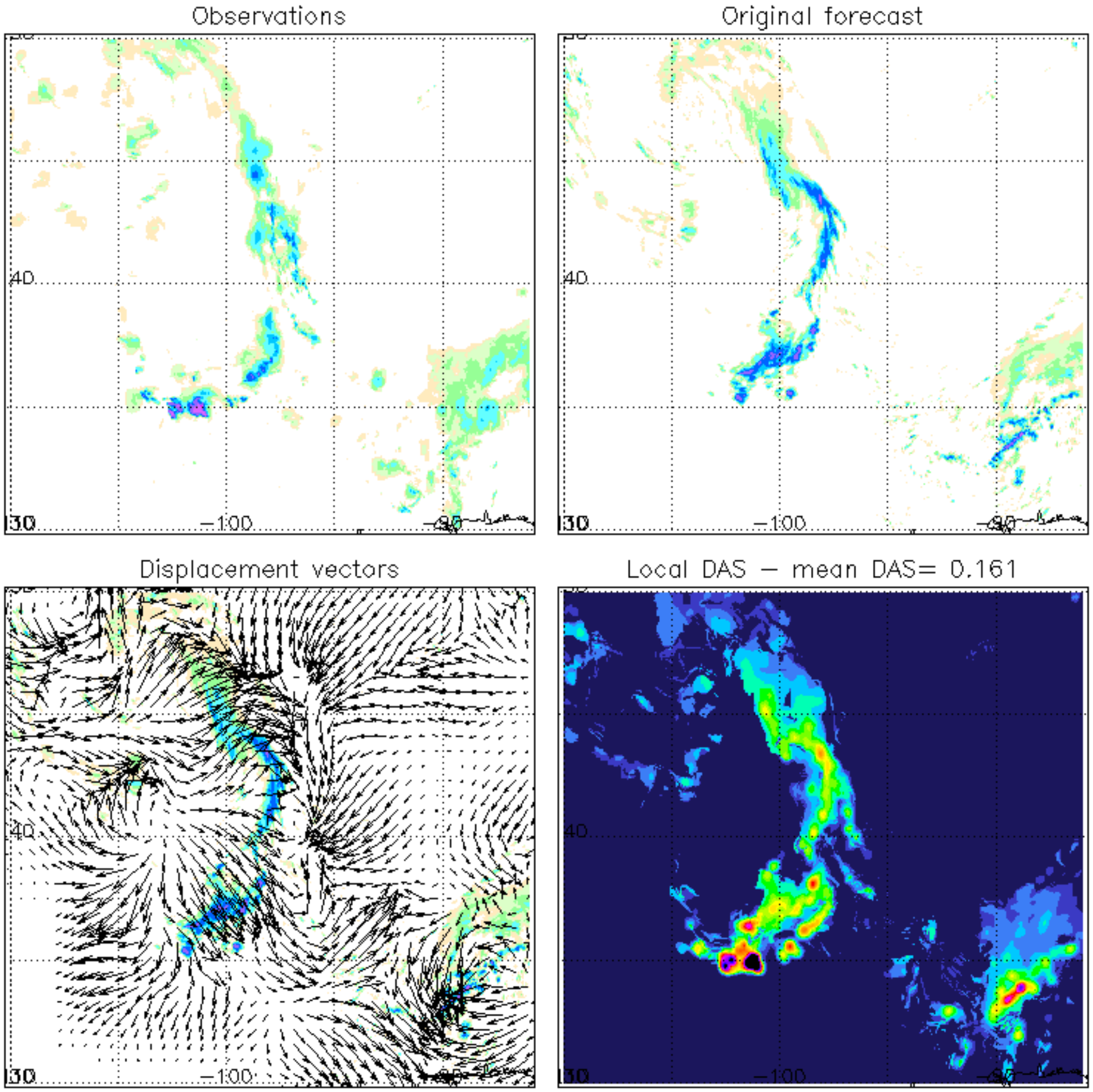
satellite | orig.model | morphed model

Example: DAS verification of precipitation forecast over USA



1. How much must the forecast be distorted in order to match the observations?
2. After morphing how much amplitude error remains in the forecast?
3. What is the overall quality of the forecast as measured by the distortion and amplitude errors together?

DAS applied to our US forecast



1. How much must the forecast be distorted in order to match the observations?
2. After morphing how much amplitude error remains in the forecast?
3. What is the overall quality of the forecast as measured by the distortion and amplitude errors together?

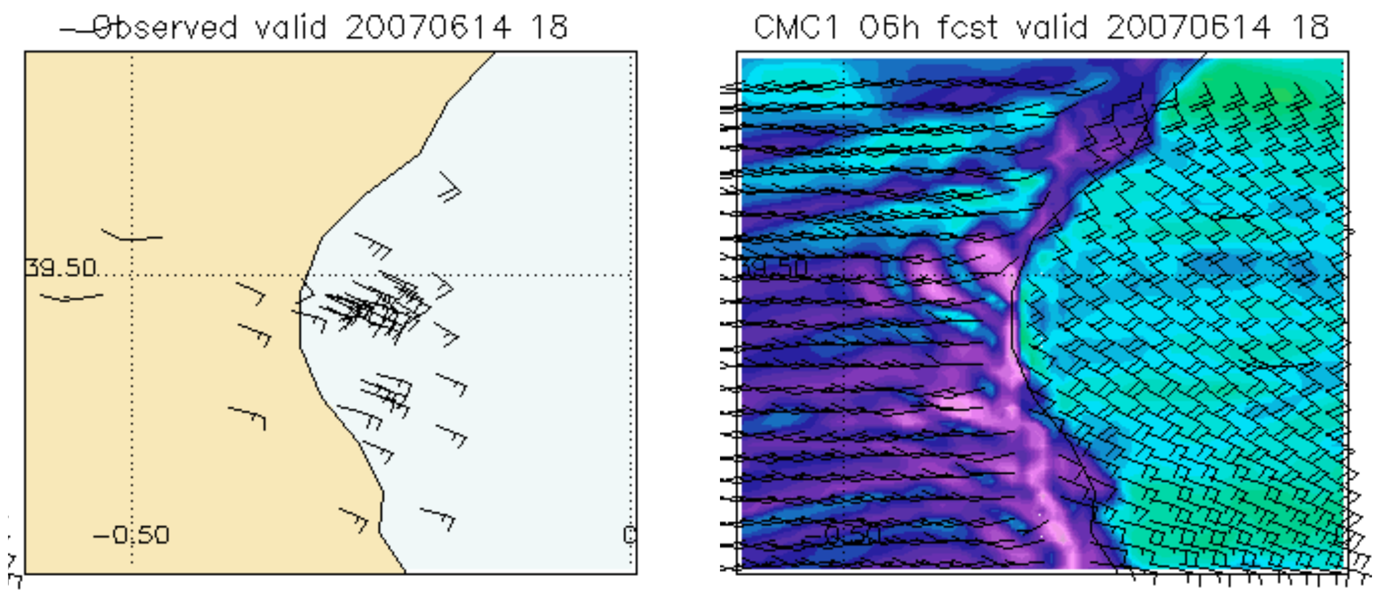
Conclusions

- What method should you use for spatial verification?
  - Depends what question(s) you would like to address

- Many spatial verification approaches
  - Neighborhood (fuzzy) – credit for "close" forecasts
  - Scale decomposition – scale-dependent error
  - Object-oriented – attributes of features
  - Field verification – phase and amplitude errors

What method(s) could you use to verify

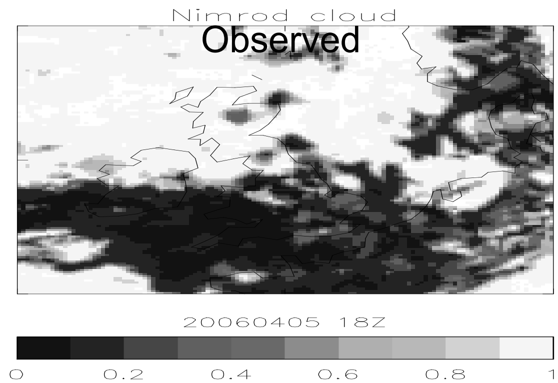
Wind forecast (sea breeze)

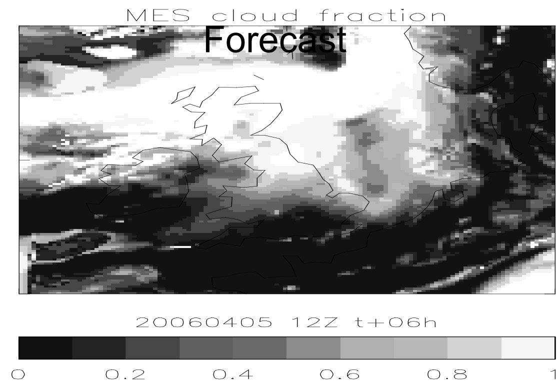


Neighborhood (fuzzy) – credit for "close" forecasts
Scale decomposition – scale-dependent error
Object-oriented – attributes of features
Field verification – phase and amplitude errors

What method(s) could you use to verify

Cloud forecast

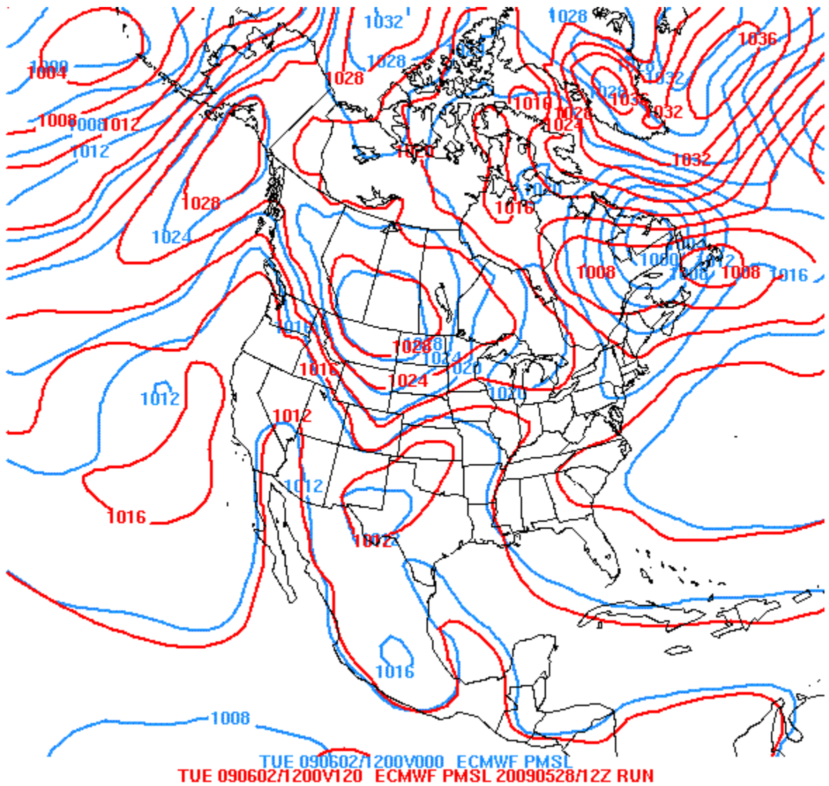




Neighborhood (fuzzy) – credit for "close" forecasts
Scale decomposition – scale-dependent error
Object-oriented – attributes of features
Field verification – phase and amplitude errors

What method(s) could you use to verify

Mean sea level pressure forecast



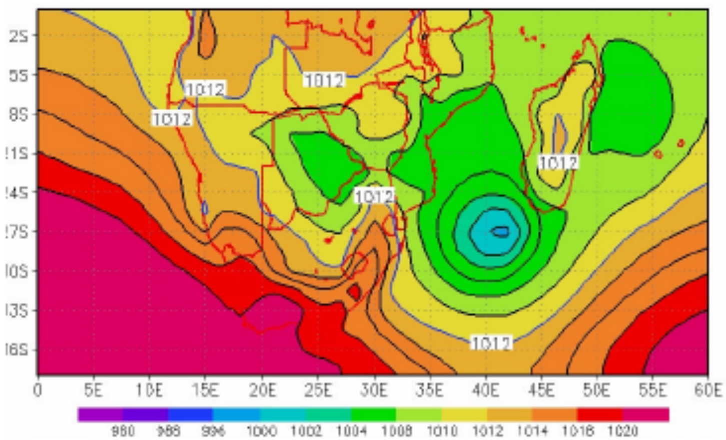
5-day forecast
Analysis

Neighborhood (fuzzy) – credit for "close" forecasts
Scale decomposition – scale-dependent error
Object-oriented – attributes of features
Field verification – phase and amplitude errors

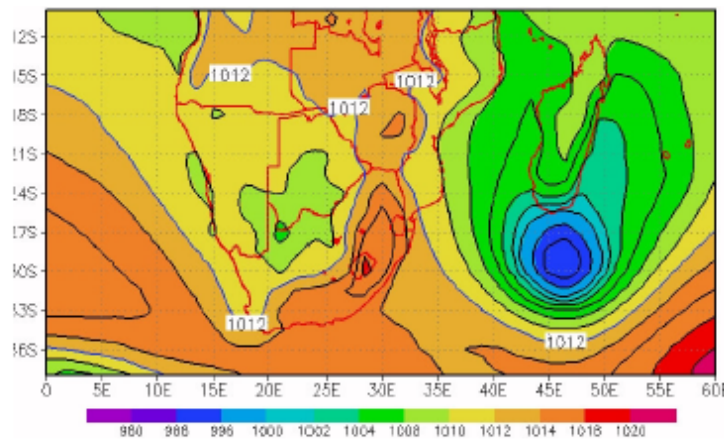
What method(s) could you use to verify

Tropical cyclone forecast

Observed



3-day forecast



Neighborhood (fuzzy) – credit for "close" forecasts
Scale decomposition – scale-dependent error
Object-oriented – attributes of features
Field verification – phase and amplitude errors

That's it!