Methods for evaluating spatial fields

Barbara Brown

Joint Numerical Testbed
Research Applications Laboratory, NCAR
27 October 2010

Acknowledgements: Beth Ebert, Eric Gilleland, David Ahijevych, Mike Baldwin, Barbara Casati

Goal

Describe new methods for evaluation of spatial fields

- Many methods have been developed in the context of high resolution precipitation forecasts
- Methods have applicability in multiple other areas
  - Other parameters (e.g., wind, cloud)
  - Regional climate forecasts
  - Satellite precipitation estimates
  - Other satellite estimates

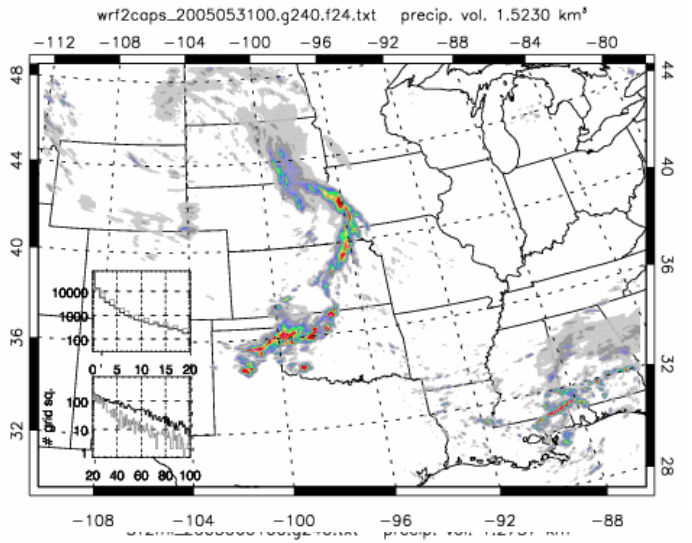
Outline

- Motivation for alternative methods
- Taxonomy of new approaches
- Comparison of capabilities
- Example applications
- Extensions

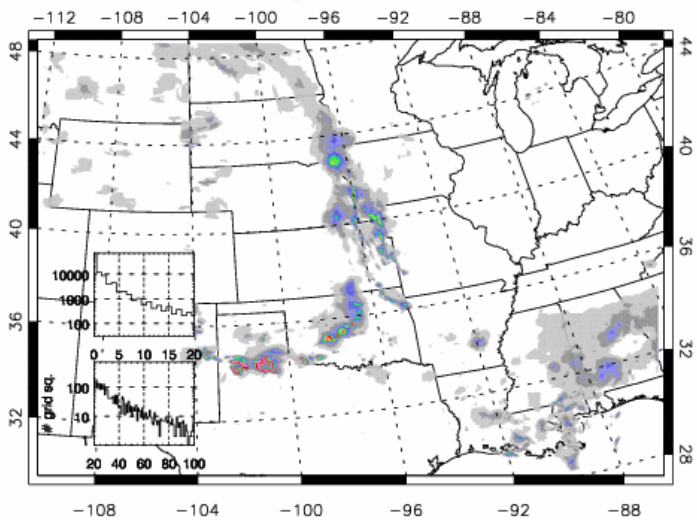
Spatial fields

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

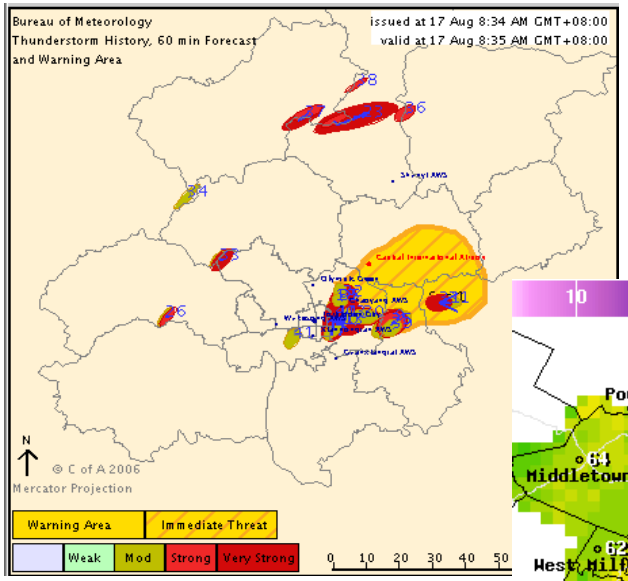
WRF model

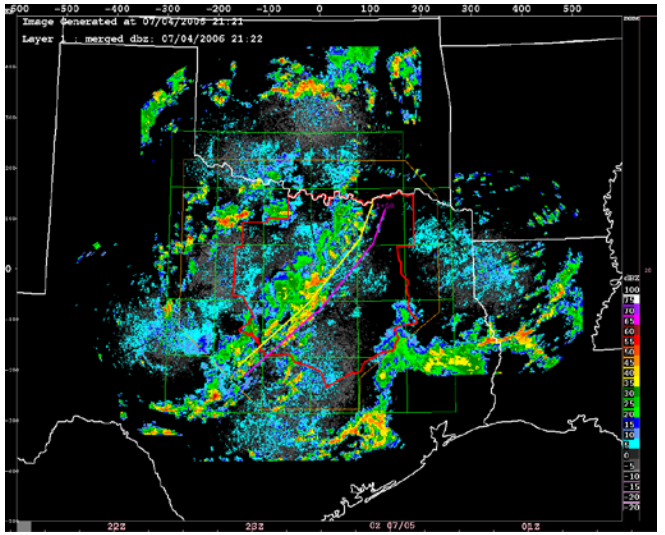


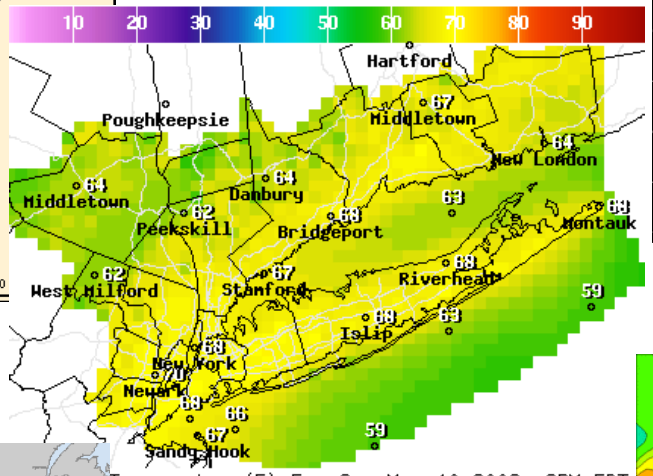
Stage II radar

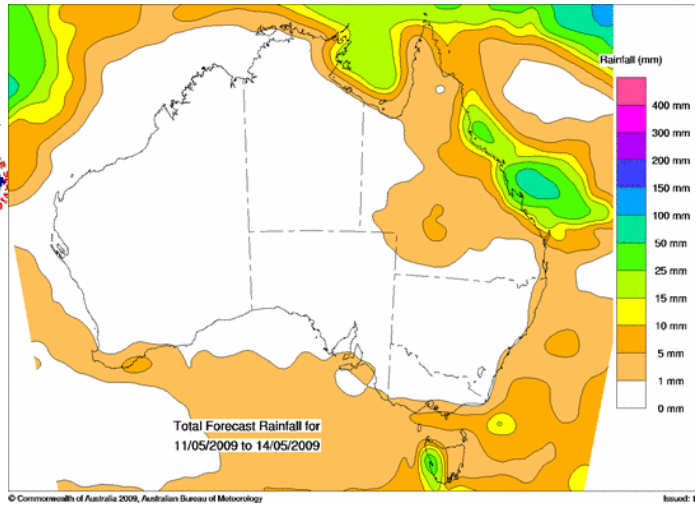


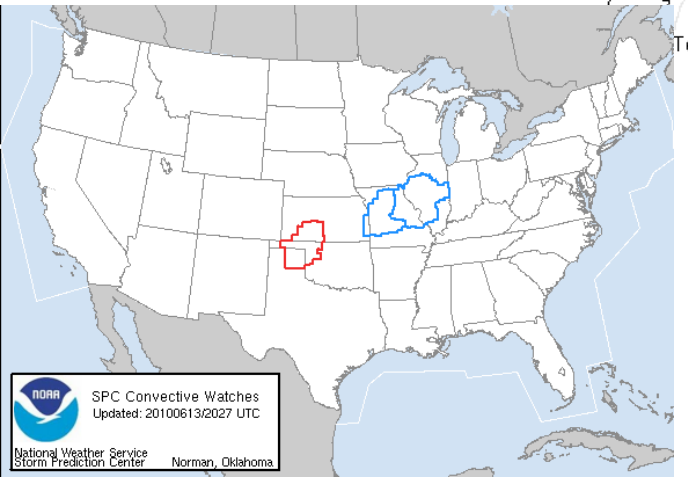
Spatial fields have many flavors (forms and scales)









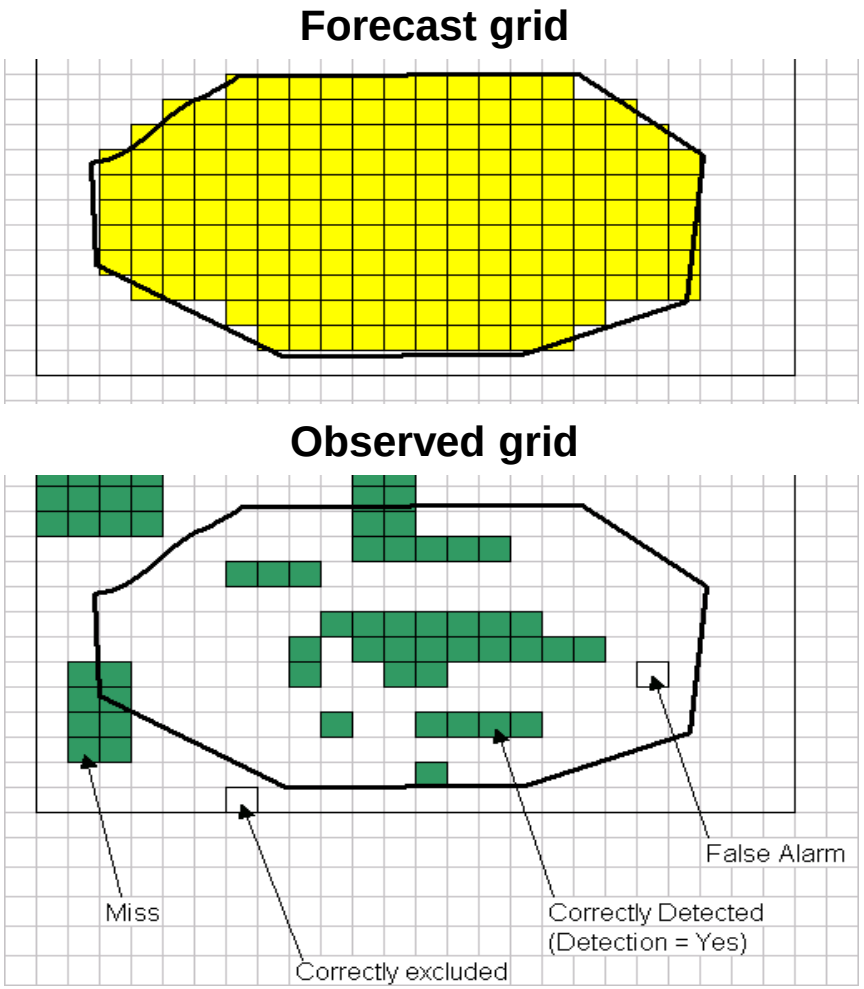


Matching two fields (forecasts and observations)

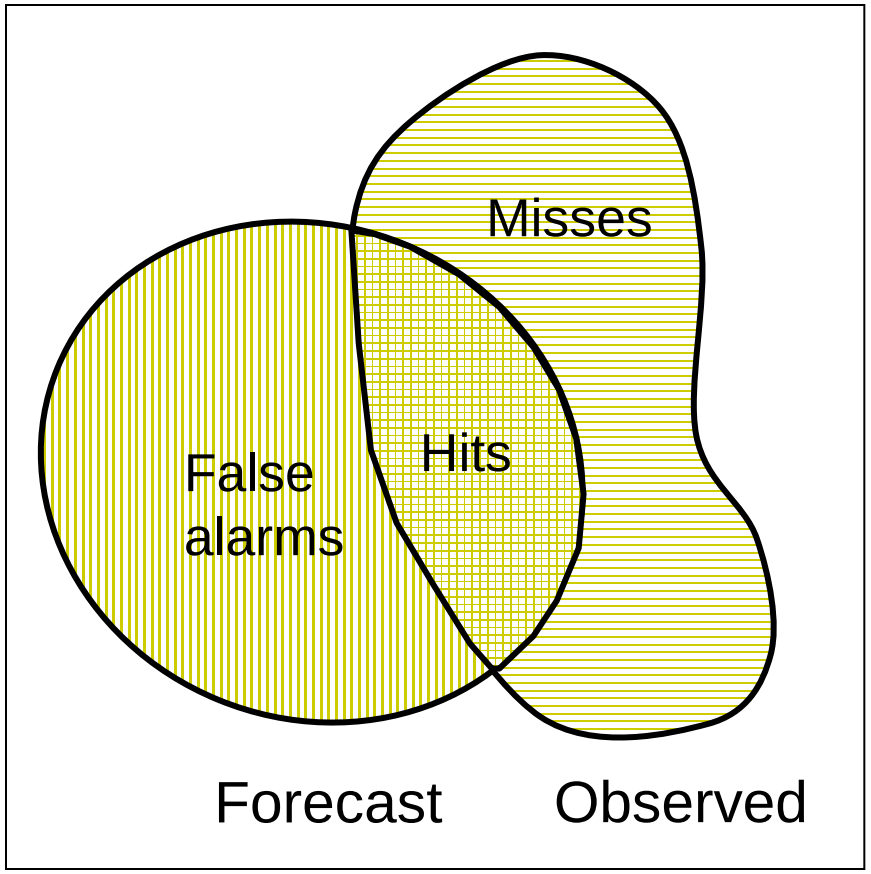
Focus: Gridded fields

Traditional grid to grid approach:

- Overlay forecast and observed grids
- Match each forecast and observation gridpoint



Traditional spatial verification using categorical scores



Contingency Table

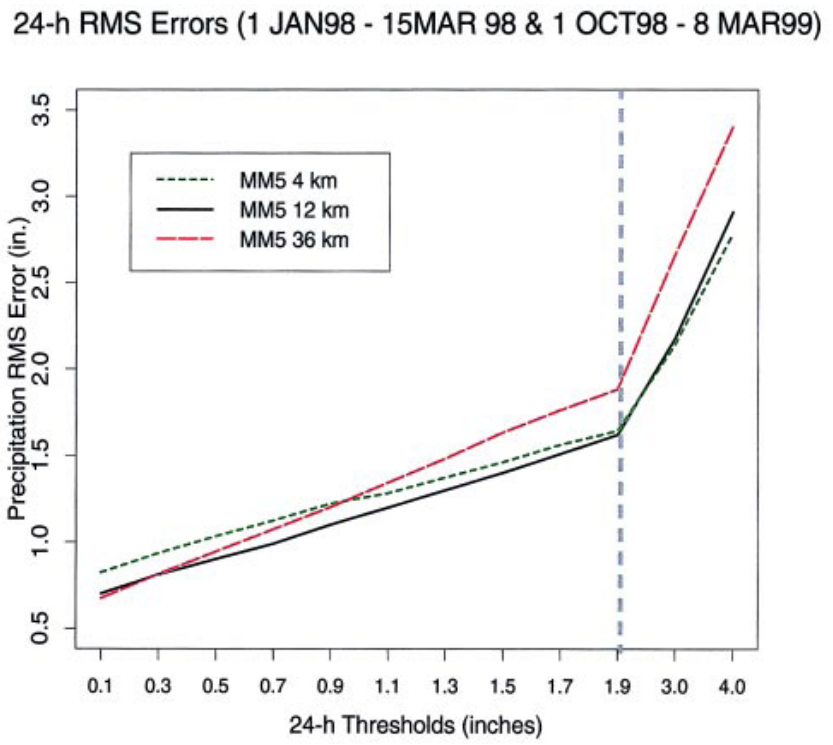
Forecast \ Observed	yes	no
yes	*hits*	*false alarms*
no	*misses*	*correct negatives*

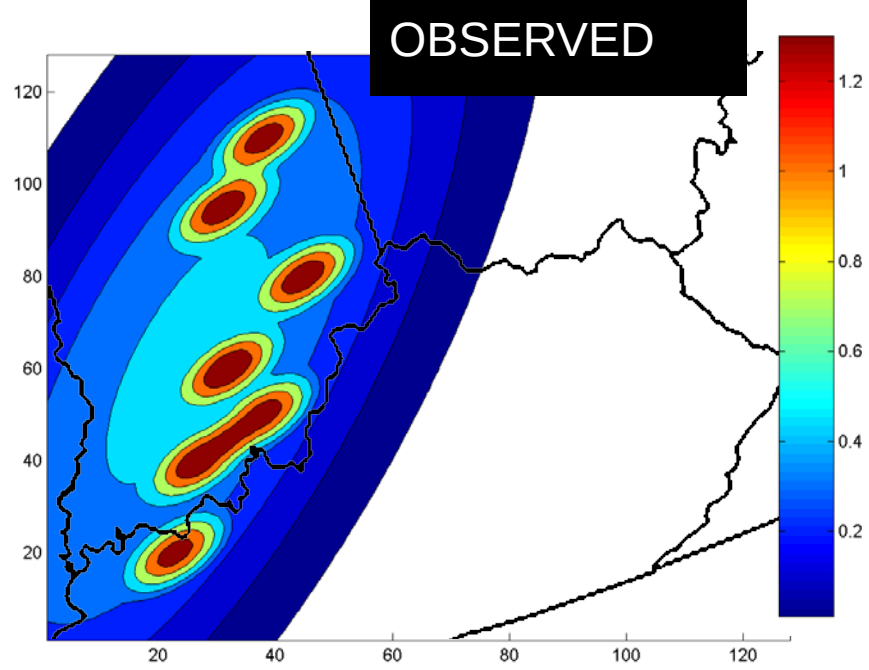
Compute

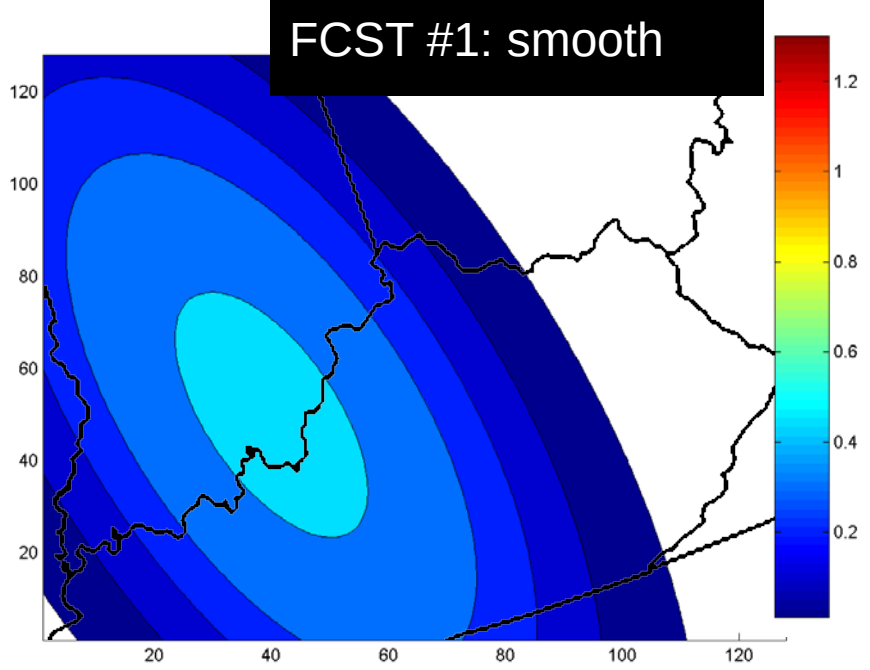
- *Contingency table statistics*: POD, FAR, Freq. Bias, CSI, GSS (= ETS)
- *Measures for continuous variables*: MSE, MAE, ME

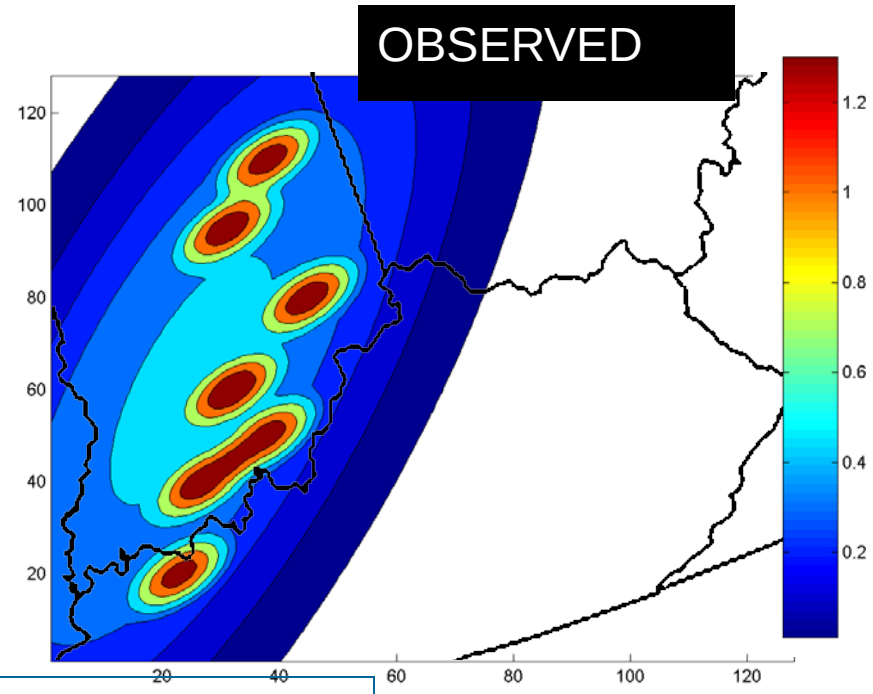
Mass et al. (2002):

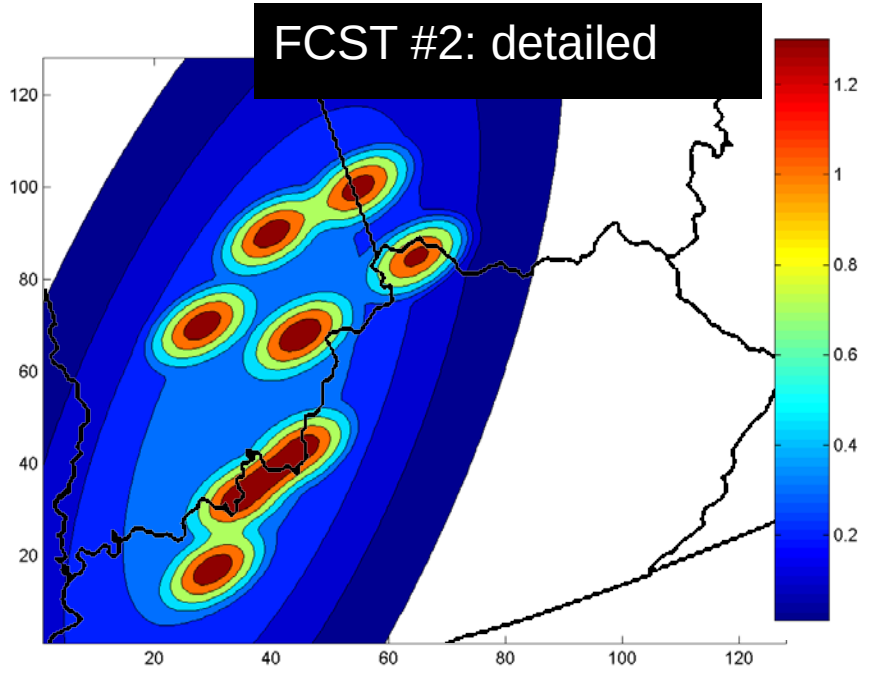
"Decreasing grid spacing in mesoscale models to less than 10–15 km generally improves the realism of the results but does not necessarily significantly improve the objectively scored accuracy of the forecasts."











From Baldwin 2002

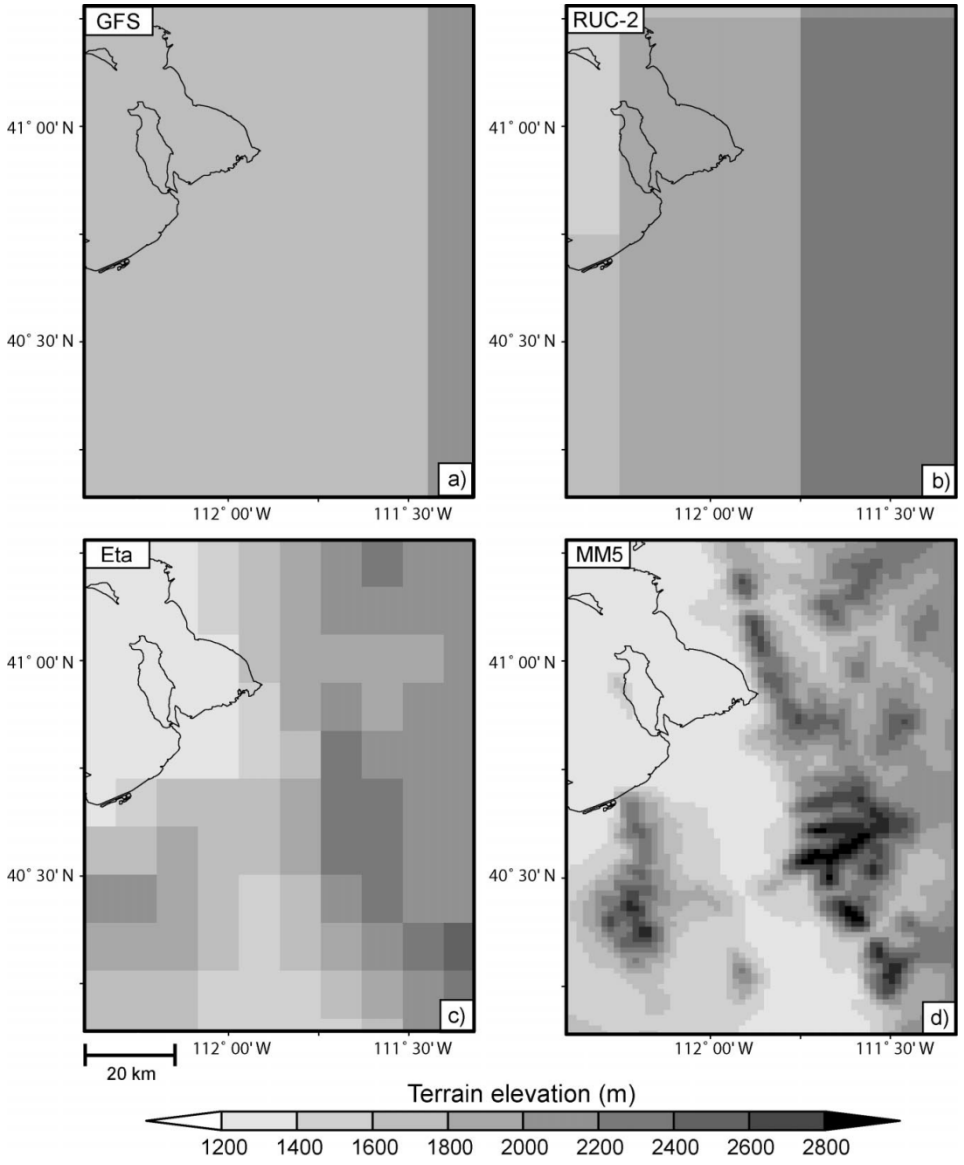
"Measures-oriented" approach to evaluating these forecasts

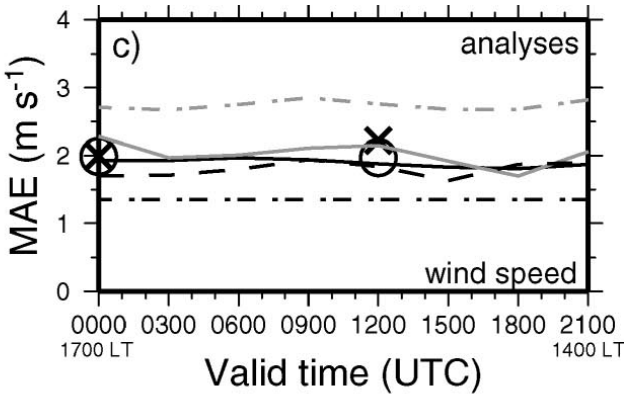
Verification Measure	Forecast #1 (smooth)	Forecast #2 (detailed)
Mean absolute error	*0.157*	0.159
RMS error	*0.254*	0.309
Bias	0.98	0.98
CSI (>0.45)	*0.214*	0.161
GSS (>0.45)	*0.170*	0.102

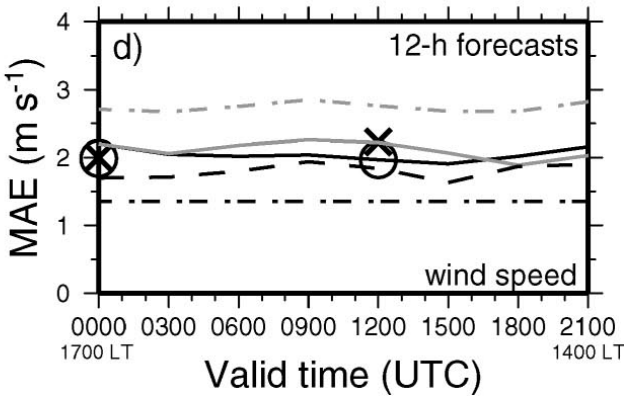
From Baldwin 2002

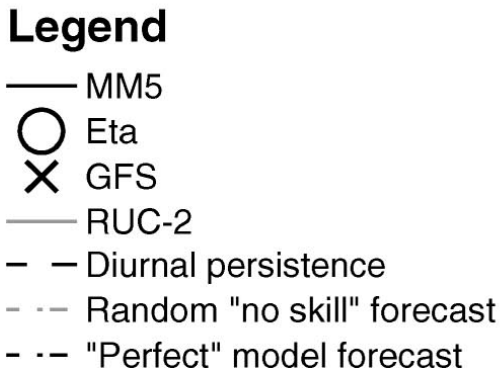
Rife et al. (2004)

"…Even though these models had horizontal grid increments that ranged over almost two orders of magnitude, the highest resolution MM5 with a 1.33-km grid increment exhibited a forecast performance similar to that of the other models in terms of grid-average, conventional verification metrics. This is in spite of the fact that the MM5 is the only model capable of reasonably representing the complex terrain of the Salt Lake City region that exerts a strong influence on the local circulation patterns."









Rife et al. 2004

What are the issues with the traditional approaches?

- "Double penalty" problem
- Scores may be insensitive to the *size* of the errors or the *kind* of errors
- Small errors can lead to very poor scores
- Forecasts are generally rewarded for being smooth
- Verification measures don't provide
  - Information about kinds of errors (Placement? Intensity? Pattern?)
  - Diagnostic information
    - 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?

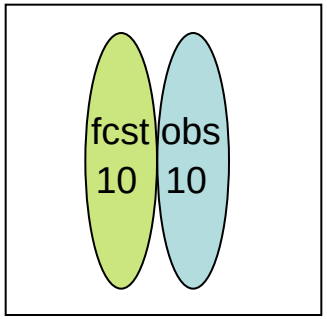
Double penalty problem

Traditional approach requires an ***exact match*** between forecasts and observations at every grid point to score a hit

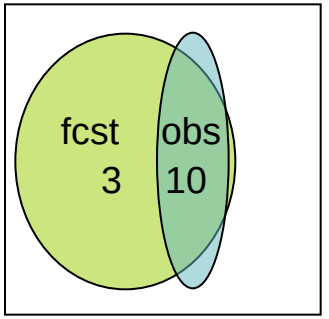
Double penalty:

(1) Event predicted where it did not occur => ***False alarm***

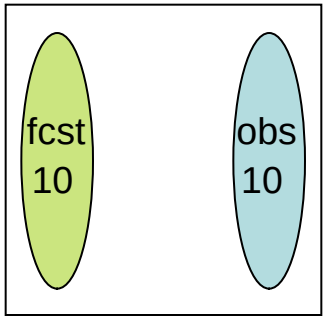
(2) No event predicted where it did occur => ***Miss***



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



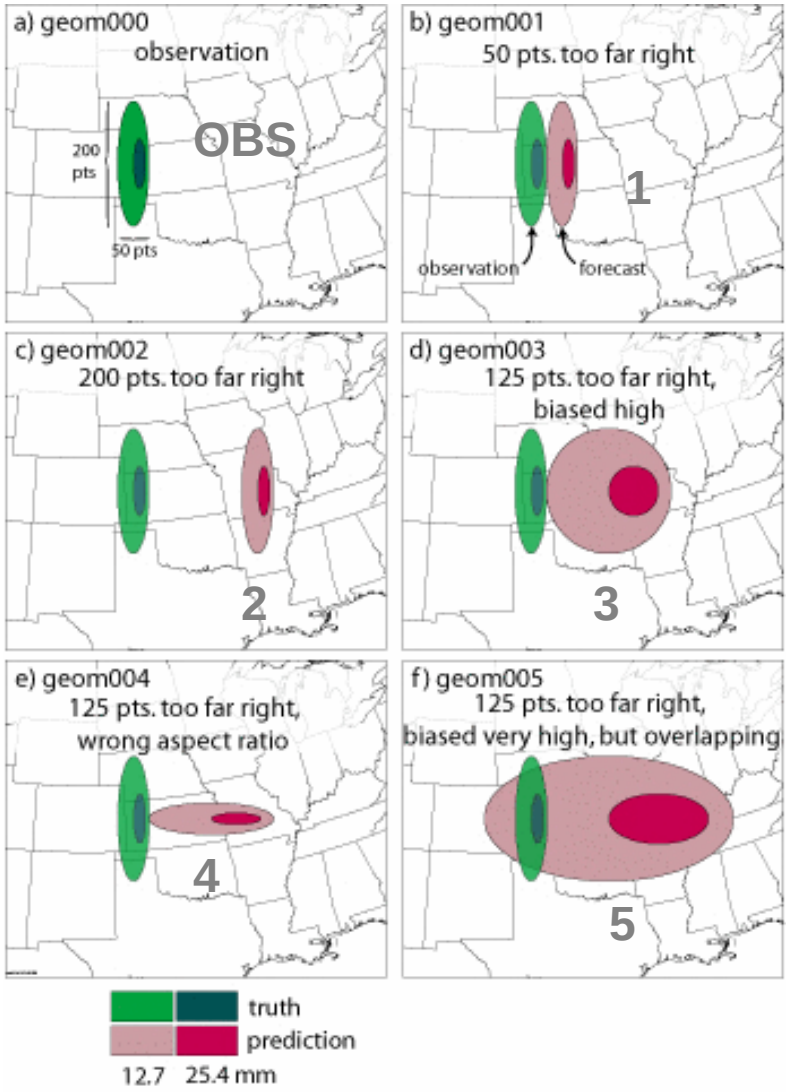
Summary: What are the issues with the traditional approaches?

- "Double penalty" problem
- Scores may be insensitive to the *size* of the errors or the *kind* of errors
- Small errors can lead to very poor scores
- Forecasts are generally rewarded for being smooth
- Verification measures don't provide
  - Information about kinds of errors (Placement? Intensity? Pattern?)
  - Diagnostic information
    - 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?

Traditional approach

Consider gridded forecasts and observations of precipitation

Which is better?



Traditional approach

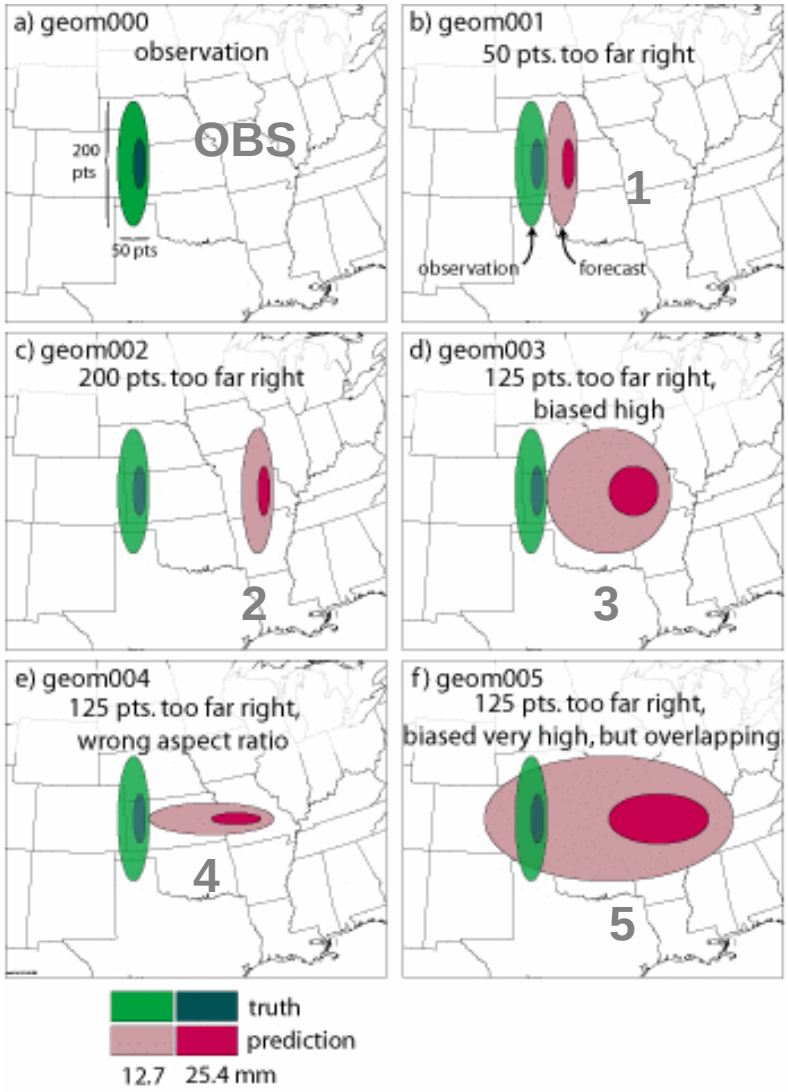
Scores for Examples 1-4:

Correlation Coefficient = -0.02
Probability of Detection = 0.00
False Alarm Ratio = 1.00
Hanssen-Kuipers = -0.03
Gilbert Skill Score (ETS) = -0.01

Scores for Example 5:

Correlation Coefficient = 0.2
Probability of Detection = 0.88
False Alarm Ratio = 0.89
Hanssen-Kuipers = 0.69
Gilbert Skill Score (ETS) = 0.08

Forecast 5 is "Best"



Summary: What are the issues with the traditional approaches?

- "Double penalty" problem
- Scores may be insensitive to the *size* of the errors or the *kind* of errors
- Small errors can lead to very poor scores
- Forecasts are generally rewarded for being smooth
- Verification measures don't provide
  - Information about kinds of errors (Placement? Intensity? Pattern?)
  - Diagnostic information
    - 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?

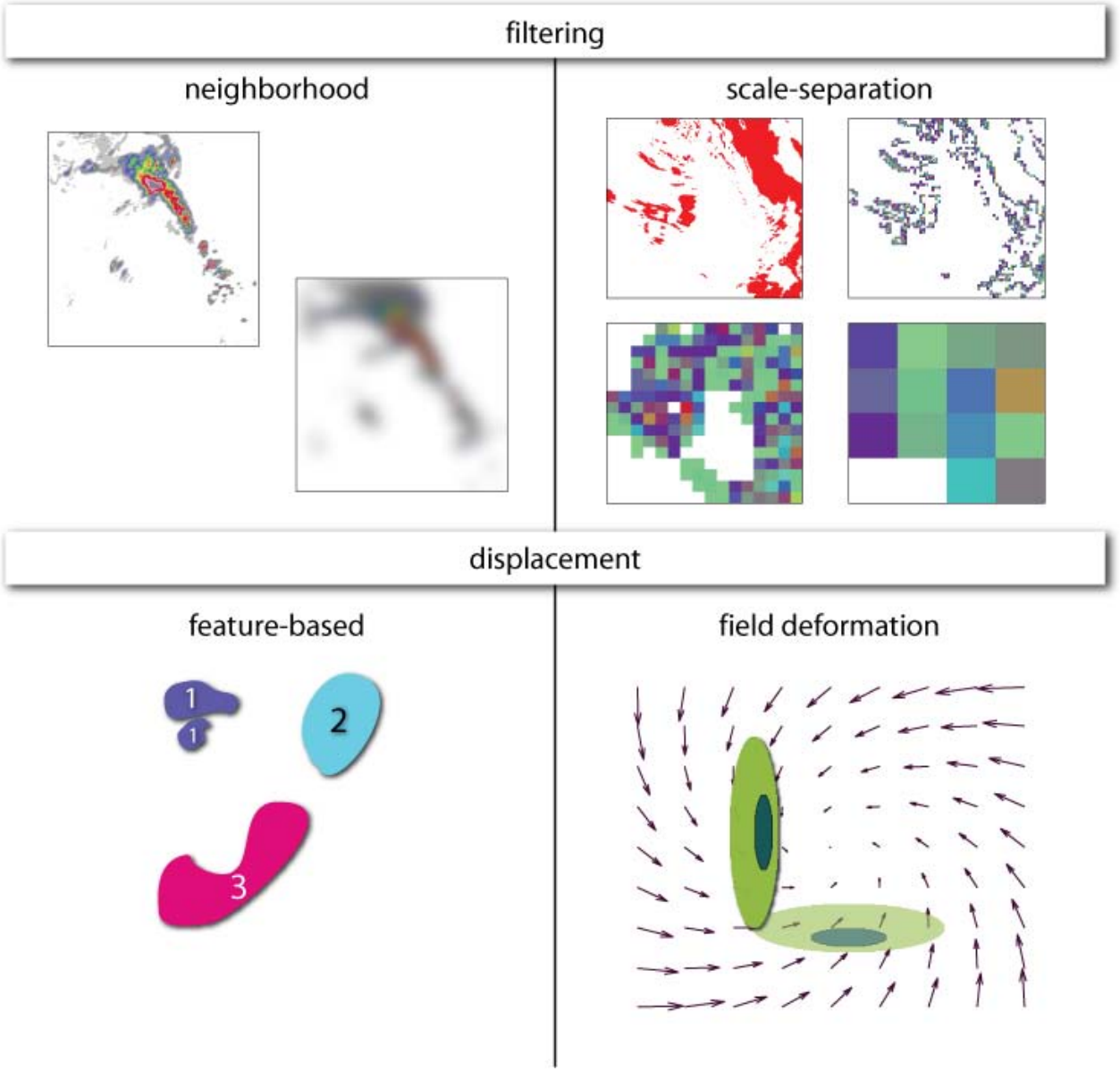
Rhetorical question:
What does CSI = 0.21 mean?

Mass et al. (2002):

"It is clear that additional approaches should join the verification toolbox… Temporal or spatial shifting of model fields could be used to verify model structures. If suitable objective verification approaches can be devised it may be possible to demonstrate increased value of high-resolution NWP."

To address the issues described here, a variety of new methods have been developed

Spatial Method Categories



New spatial verification approaches

Neighborhood

Successive smoothing of forecasts/obs

Gives credit to "close" forecasts

Scale separation

Measure scale-dependent error

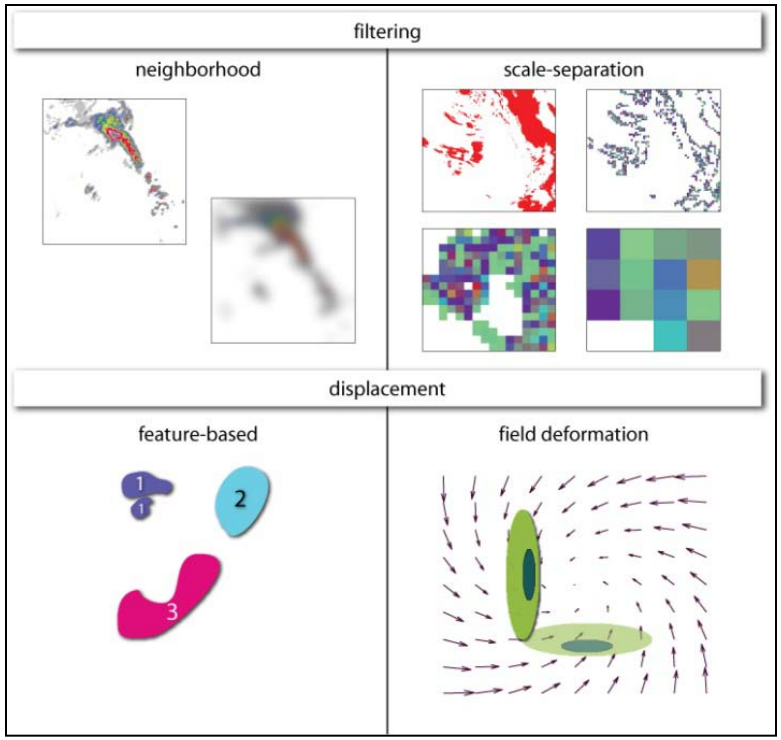
Field deformation

Measure distortion and displacement (phase error) for whole field

How should the forecast be adjusted to make the best match with the observed field?

Object- and feature-based

Evaluate attributes of identifiable features



Method Intercomparison Project (ICP)

Goals:

- Investigate and compare capabilities of new methods

  What do they tell us?
- Identify strengths and weaknesses

Activities:

- Workshops (2007,2008)
- *Weather and Forecasting* Special Collection (16 articles)

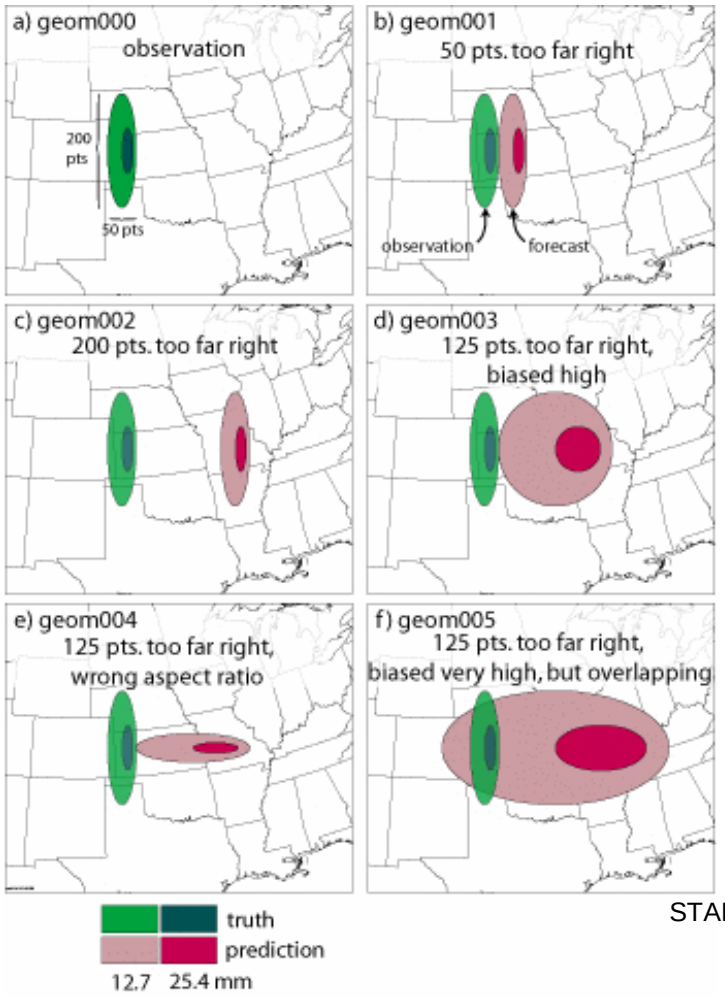
Datasets:

- Geometric cases
- Actual precipitation forecasts and obs
  - WRF precipitation forecasts (4 km)
  - Stage IV precipitation analysis
  - Resolution: 4 km
  - Domain: Central U.S.
  - Time period: May-Jun 2005 (9 focus cases)
- Perturbed cases

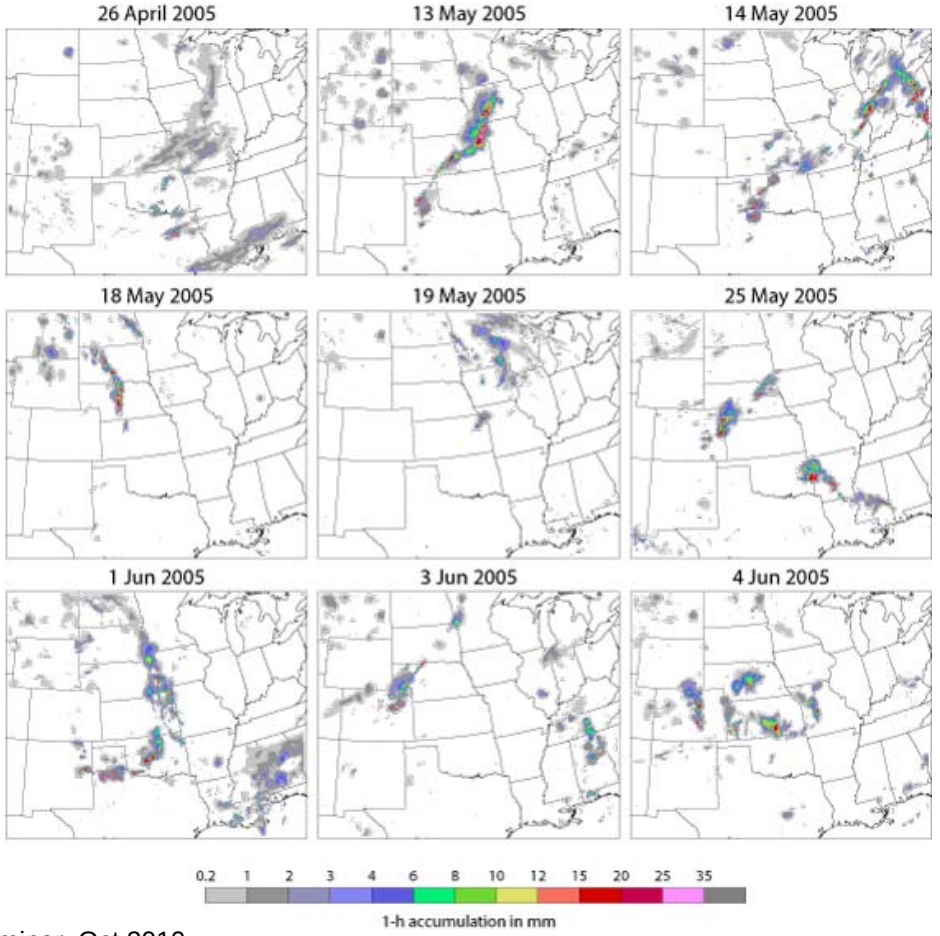
Web site: http://www.ral.ucar.edu/projects/icp/

Cases used in the ICP

Geometric cases

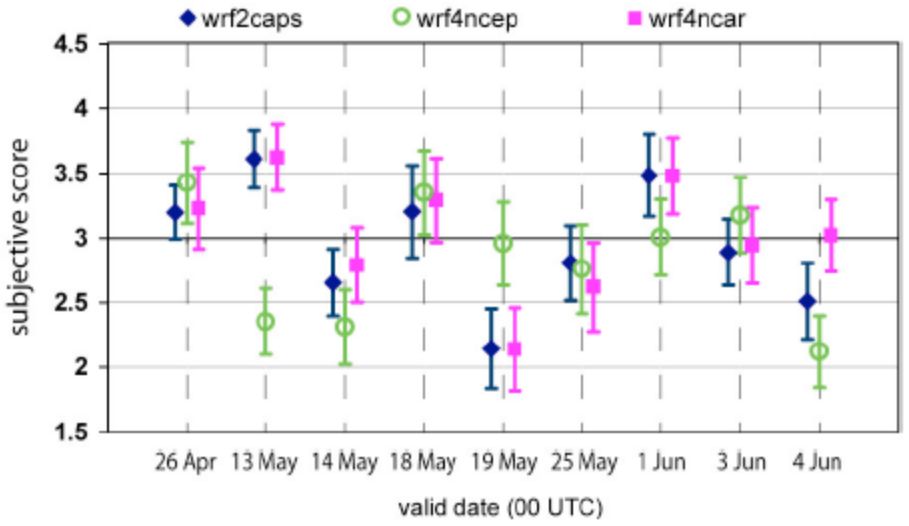


"Real" cases

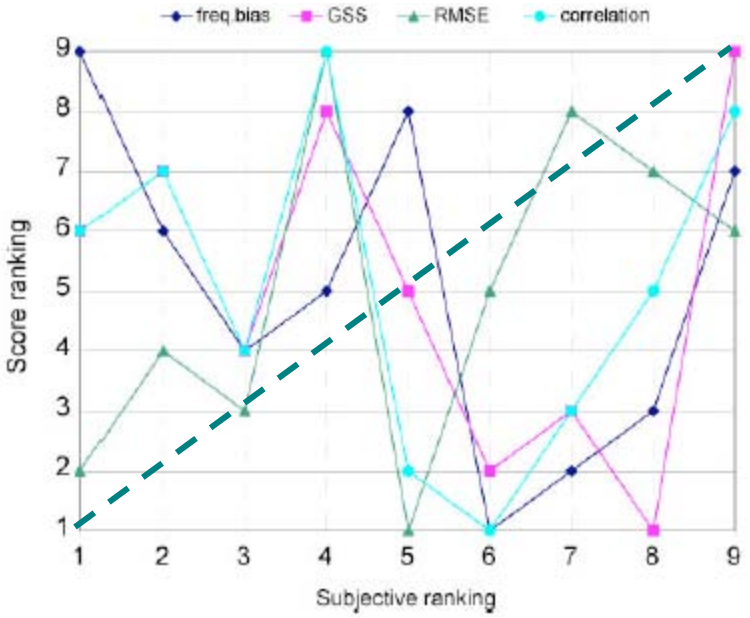


Comparison of subjective and objective evaluations

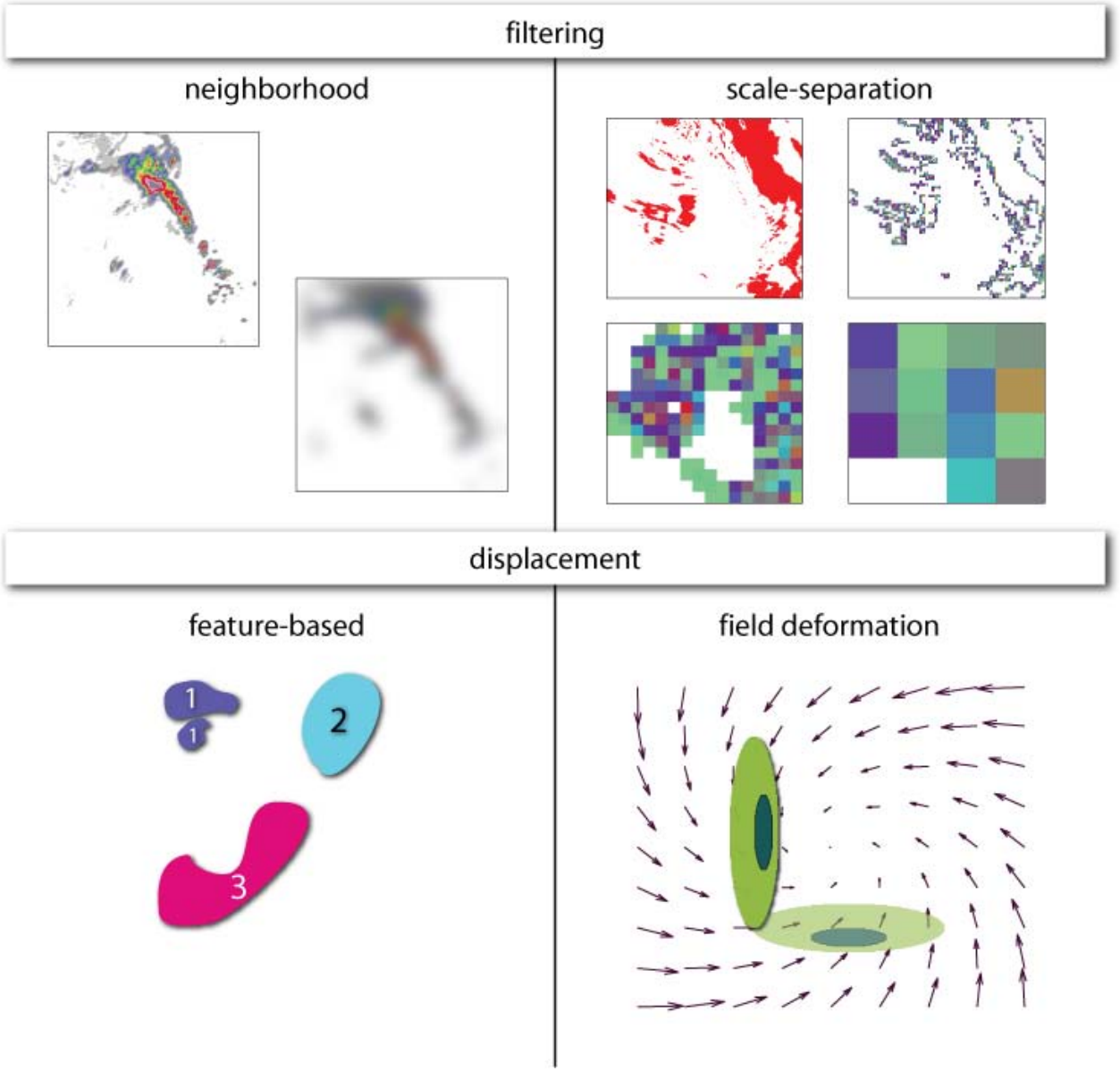
Subjective comparison



Comparison of subjective and objective "case" ranks



Spatial Method Categories



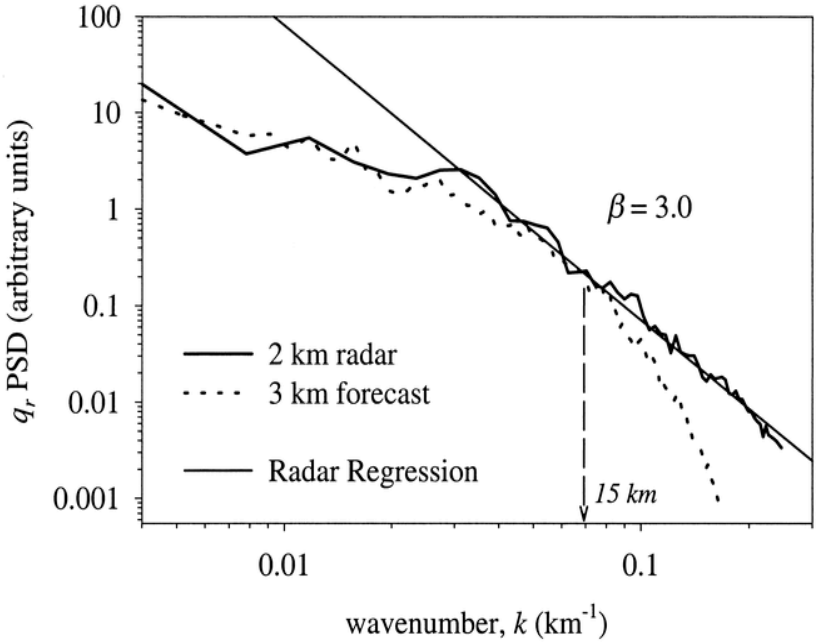
Filter Methods: Scale Separation

Scale separation methods

- Goal:

  Examine performance as a function of spatial scale

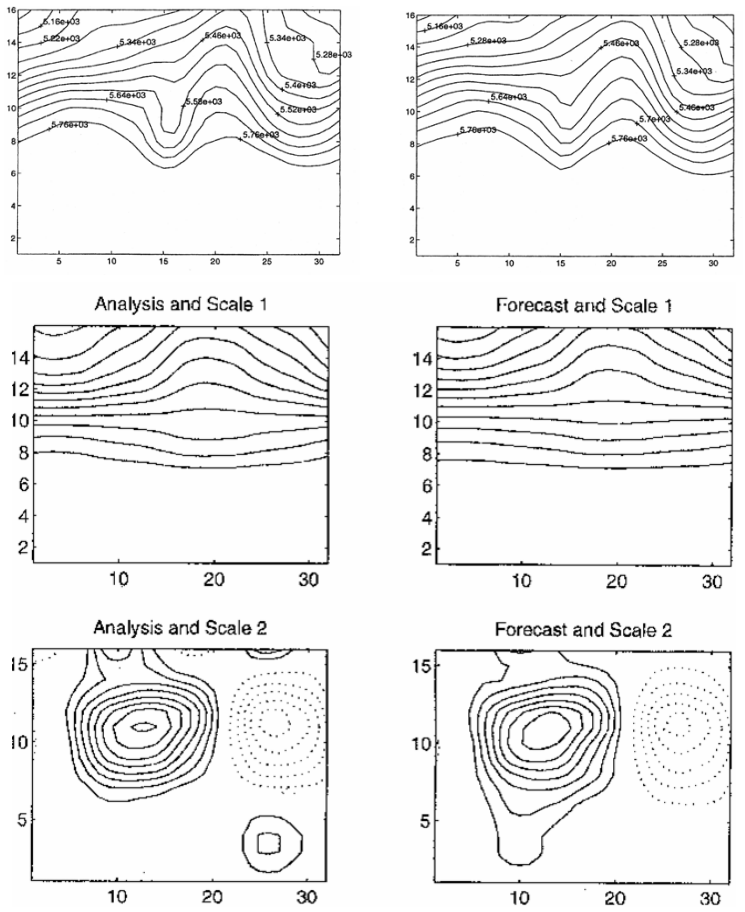
- Example: Power spectra
  - Does it look real?
  - Harris et al. (2001): compare multi-scale statistics for model and radar data



From Harris et al. 2001

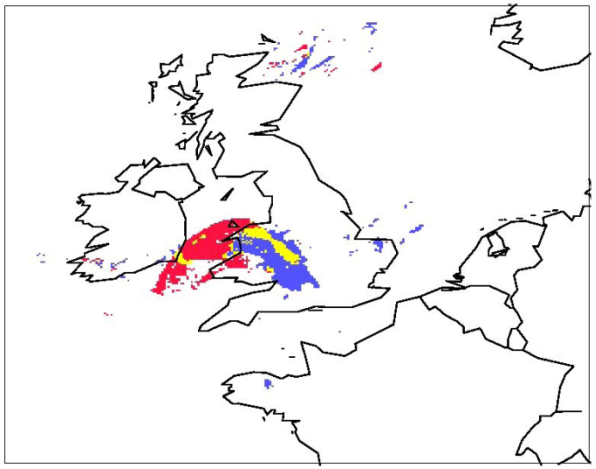
Scale decomposition

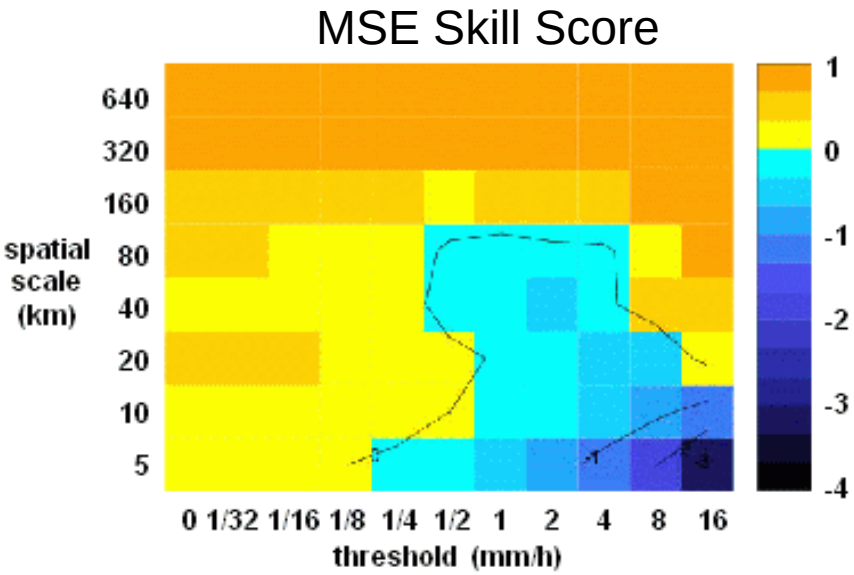
- Wavelet component analysis
  - Briggs and Levine, 1997
  - Casati et al., 2004
- Examine how different scales contribute to traditional scores



Scale separation methods

- Intensity-scale approach (Casati et al. 2004)
  - Discrete wavelet
  - Estimate performance as a function of scale
- Multi-scale variability (Zapeda-Arce *et al.* 2000; Harris *et al.* 2001 Mittermaier 2006)
- Variogram (Marzban and Sandgathe 2009)



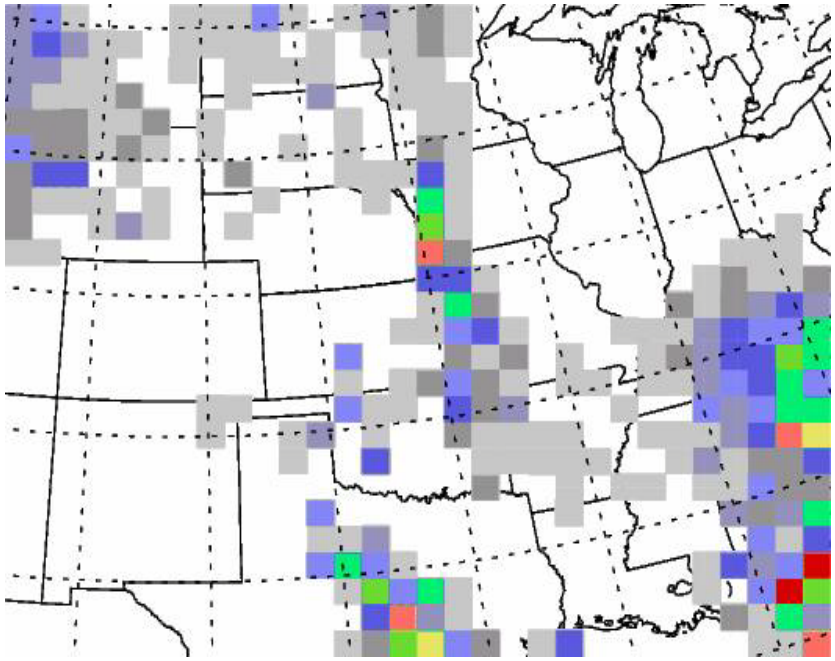


Neighborhood verification

Goal:

Examine forecast performance in a region; don't require exact matches

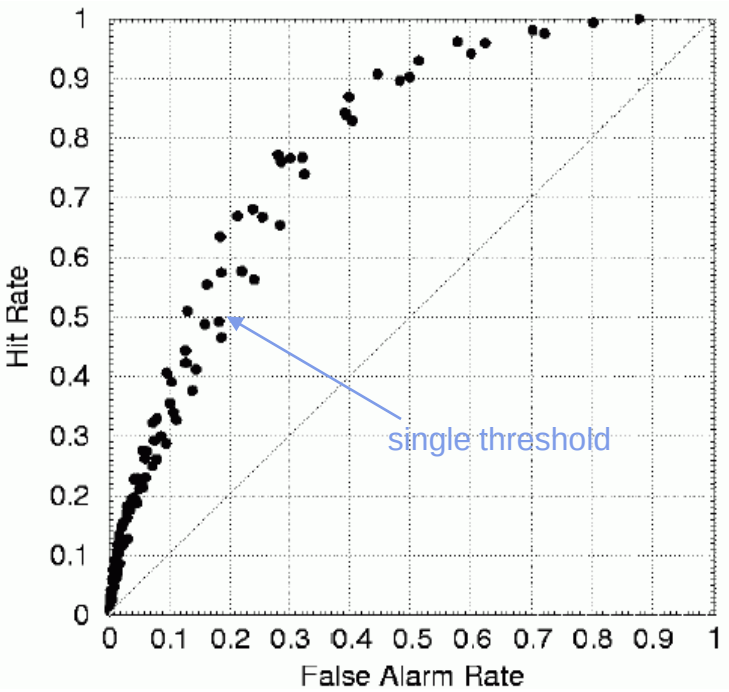
- Also called "fuzzy" verification
- Example: Upscaling
  - Put observations and/or forecast on coarser grid
  - Calculate traditional metrics
- Provide information about scales where the forecasts have skill



Neighborhood methods

Examples :

- Distribution approach (Marsigli)
- Fractions Skill Score (Roberts 2005; Roberts and Lean 2008; Mittermaier and Roberts 2009)
- Multiple approaches (Ebert 2008, 2009) (e.g., Upscaling, Multi-event cont. table, Practically perfect)

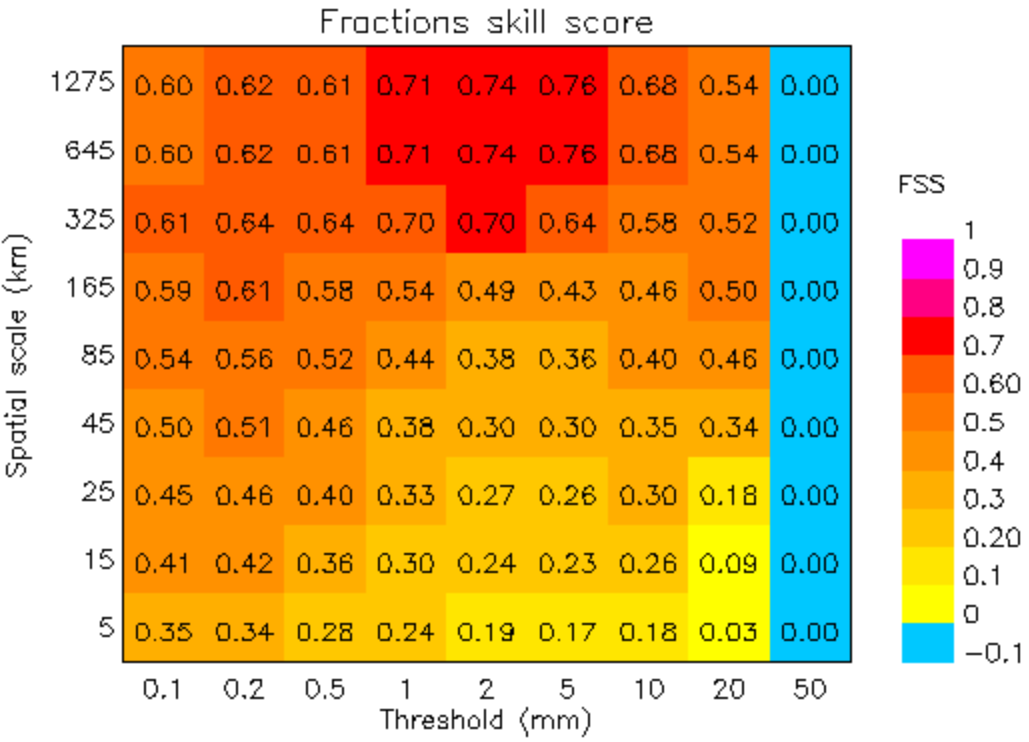


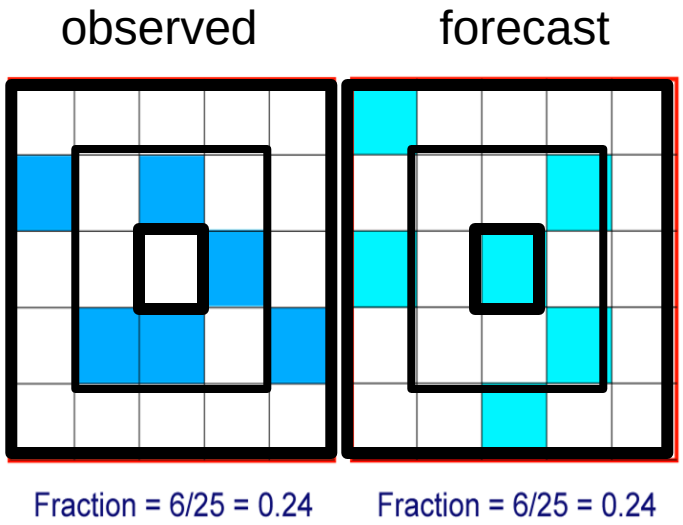
Atger, 2001

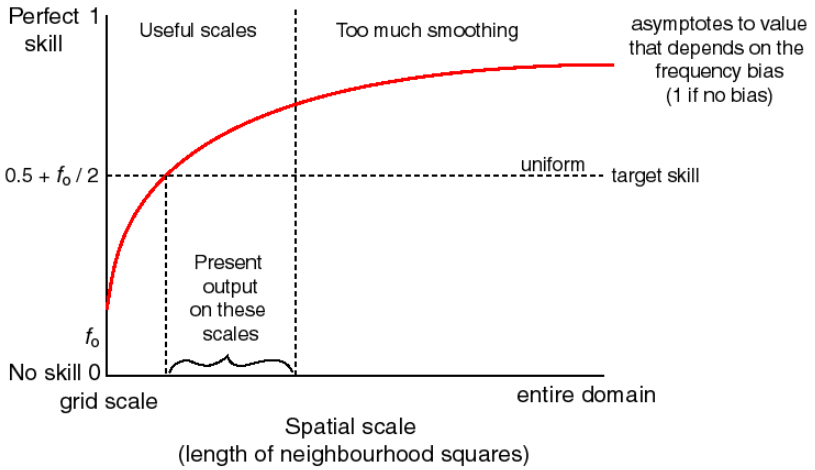
Ebert (2007; Met Applications) provides a review and synthesis of these approaches

Fractions skill score

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





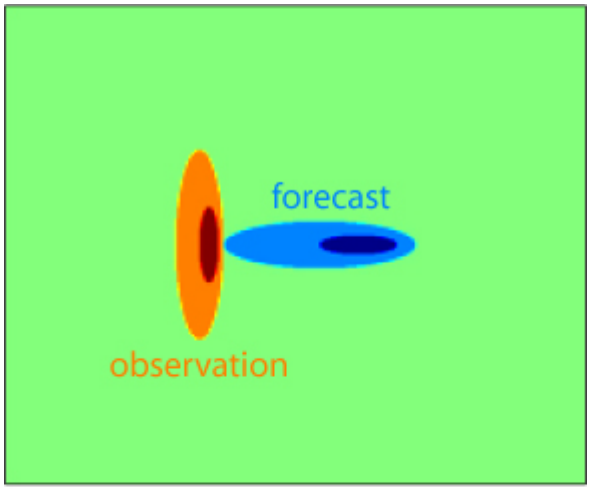


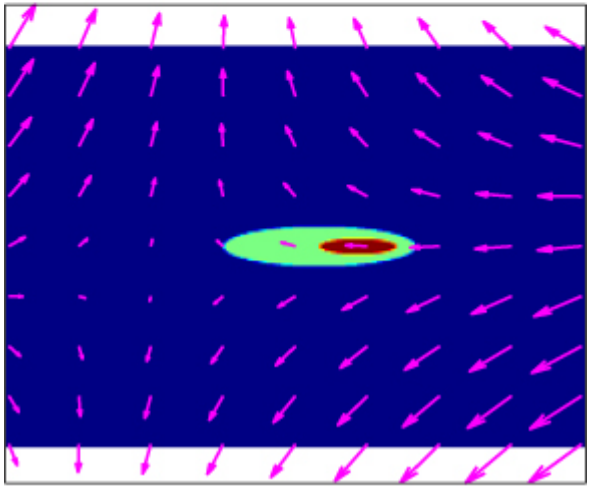
(Roberts 2005; Roberts and Lean 2007)

Field deformation

Goal:

Examine how much a forecast field needs to be transformed in order to match the observed field

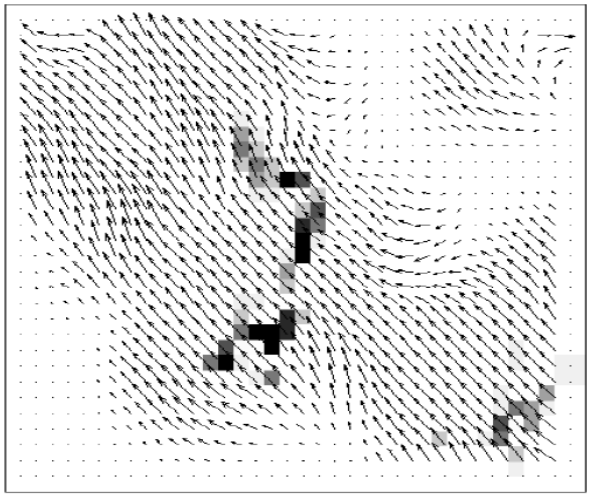




Field deformation methods

Example methods :

- Forecast Quality Index (Venugopal *et al*. 2005)
- Forecast Quality Measure/Displacement Amplitude Score (Keil and Craig 2007, 2009)
- Image Warping (Gilleland et al. 2009; Lindström *et al.* 2010; Engel 2009)

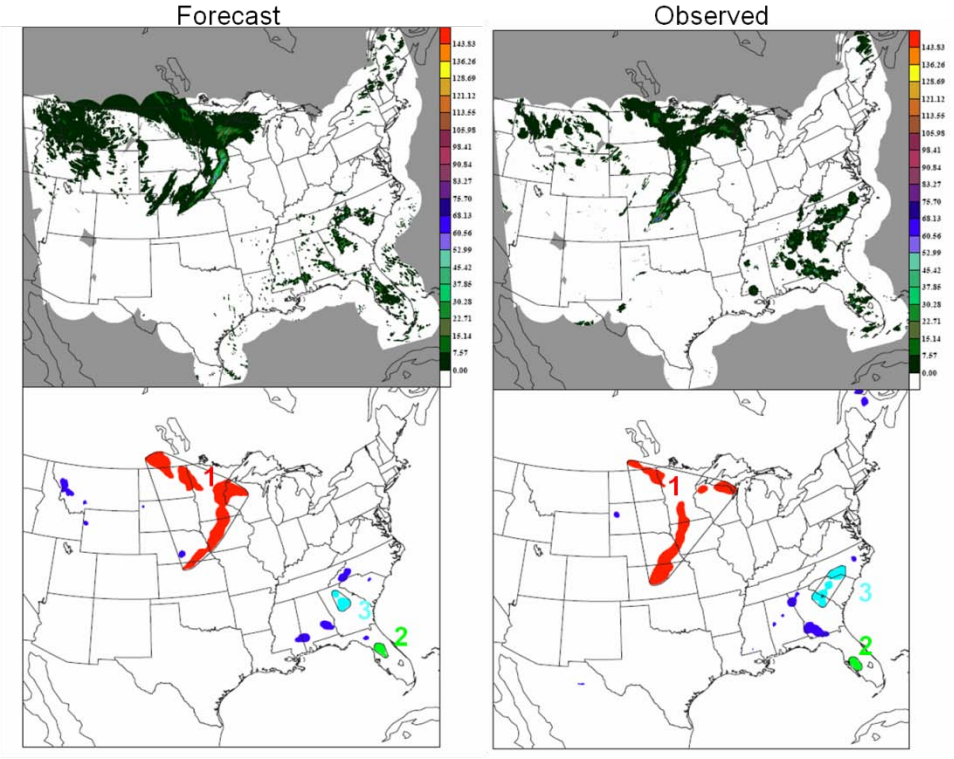


From Keil and Craig 2008

Object/Feature-based

Goals:

1. Identify relevant features in the forecast and observed fields
2. Compare attributes of the forecast and observed features



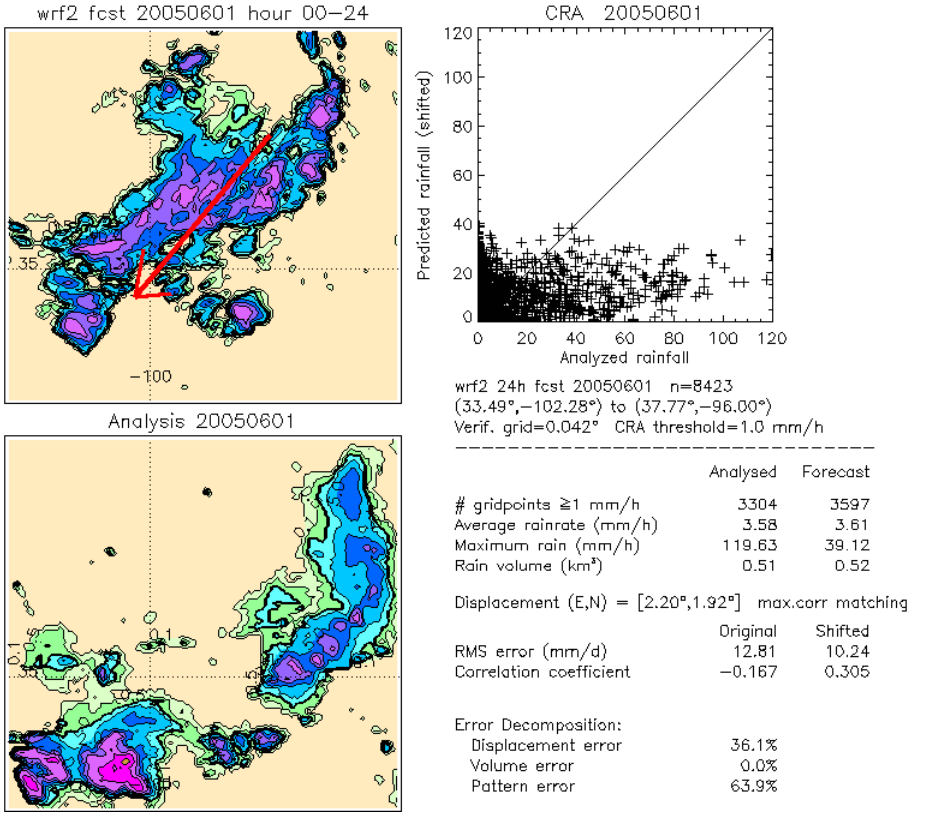
MODE example 2008

Object/Feature-based

Example methods:

- Cluster analysis (Marzban and Sandgathe 2006a,b)
- Composite (Nachamkin 2005, 2009)
- Contiguous Rain Area (CRA) (Ebert and Gallus 2009)
- Procrustes (Micheas *et al.* 2007, Lack *et al*. 2009)
- SAL (Wernli *et al.* 2008, 2009)
- MODE (Davis et al. 2005,2009)

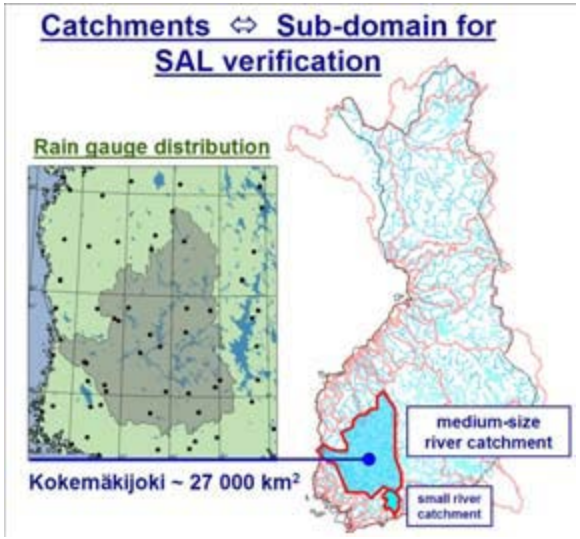
CRA example (Ebert and Gallus)



The CRA method measures displacement and *estimates error due to displacement, pattern, and volume*

Structure-Amplitude-Location (SAL)

- Focuses on features of objects (storms) in defined regions – e.g., watershed
- Goal is to characterize specific attributes of forecast performance in the watershed region



SAL Attributes

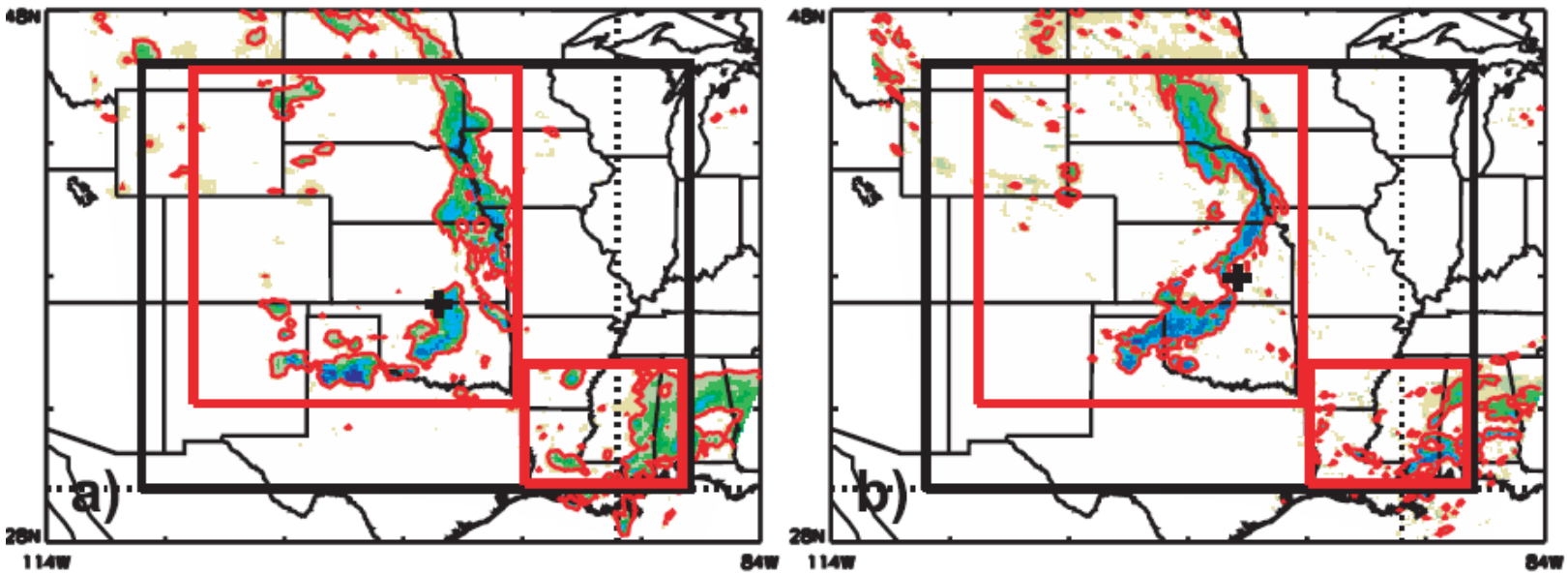
Structure

Amplitude

Location

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

SAL verification results



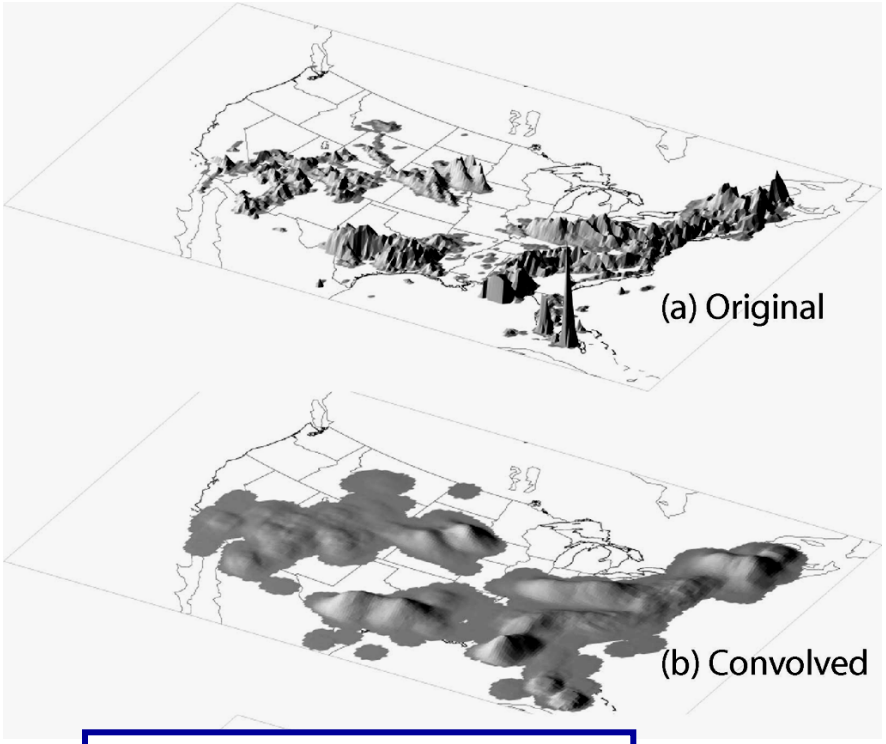
observed forecast

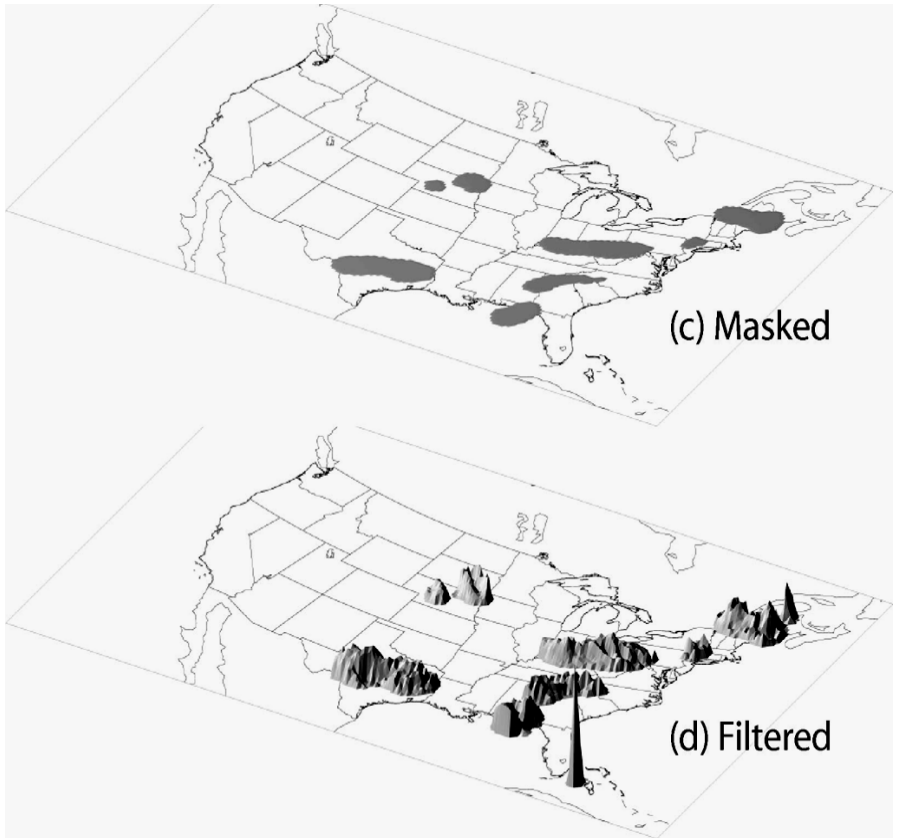
1. Is the domain average precipitation correctly forecast? A = 0.21 / 0.42
2. Is the mean location of the precipitation distribution in the domain correctly forecast? L = 0.06 / 0.08
3. Does the forecast capture the typical structure of the precipitation field (e.g., large broad objects vs. small peaked objects)?

   S = 0.46 / -1.33 (perfect=0)

MODE – Method for Object-based Diagnostic Evaluation

MODE Object identification



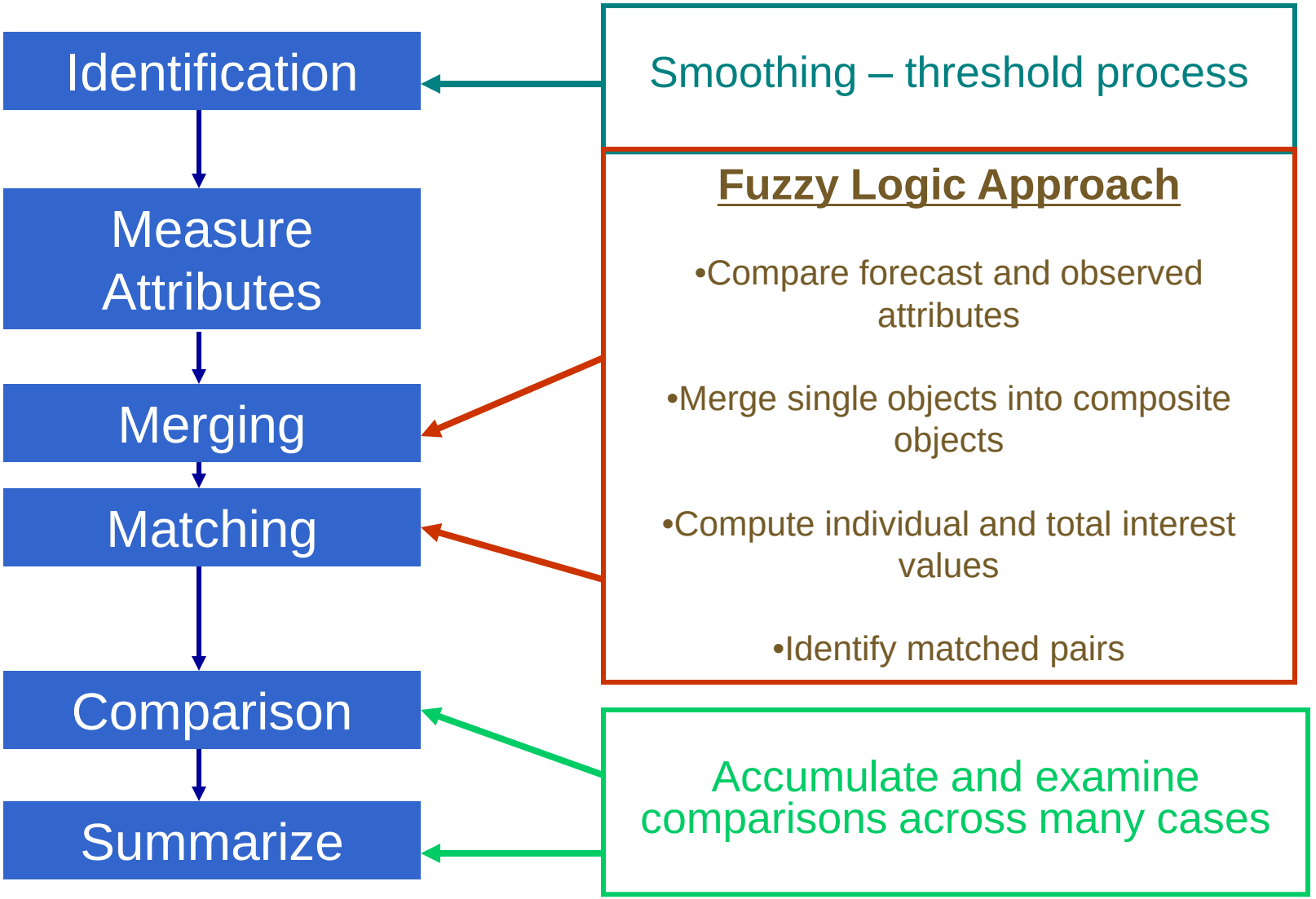


Two parameters:

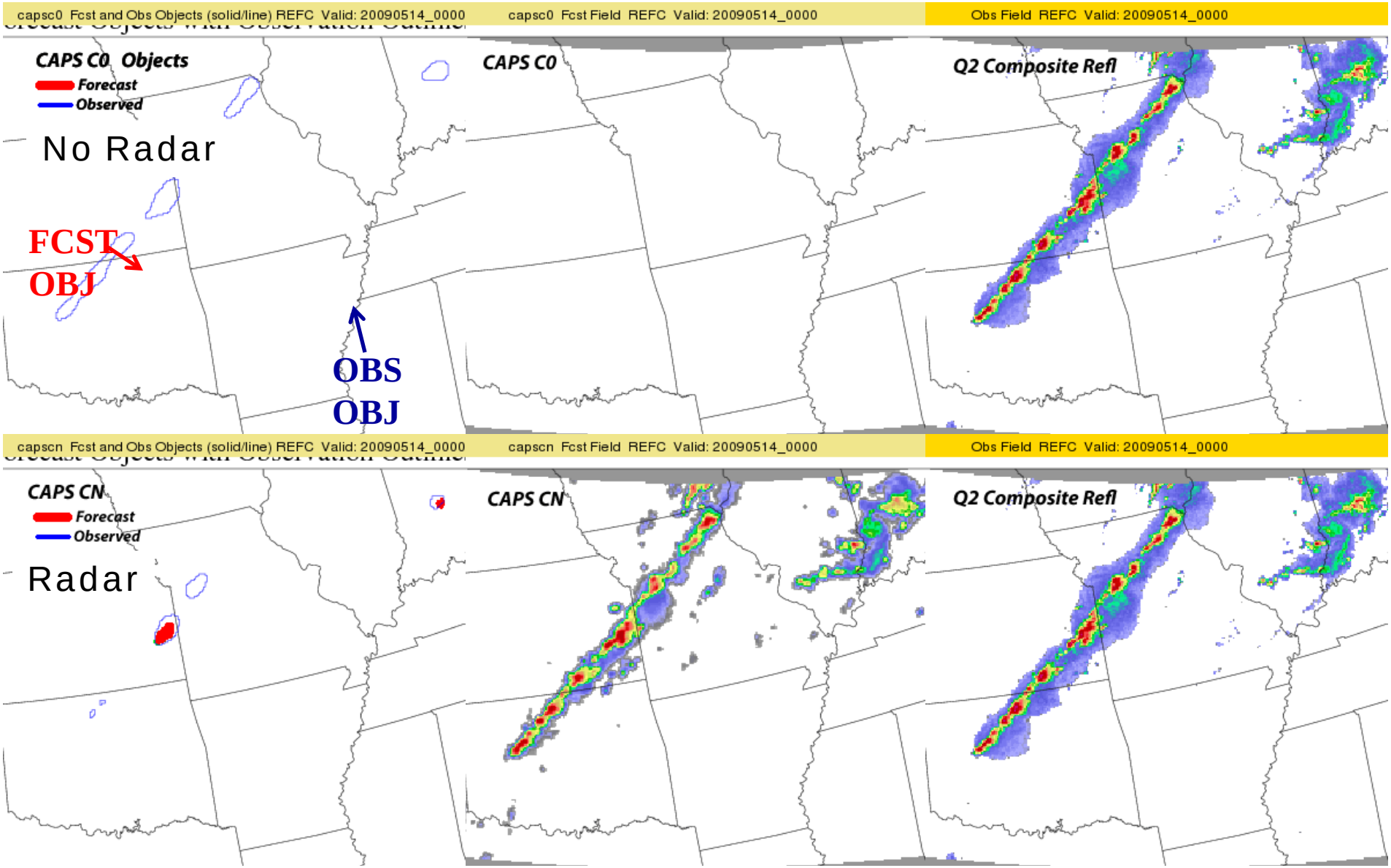
1. Convolution radius
2. Threshold

Davis et al., *MWR*, 2006; 2009

MODE methodology



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



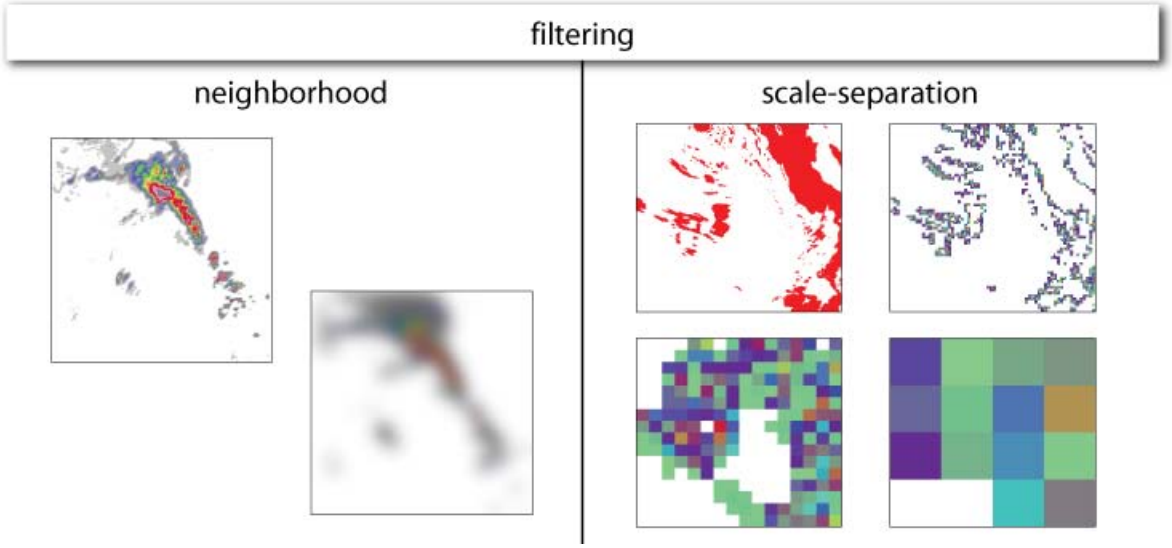
Objects

Forecast Field

Observed Field

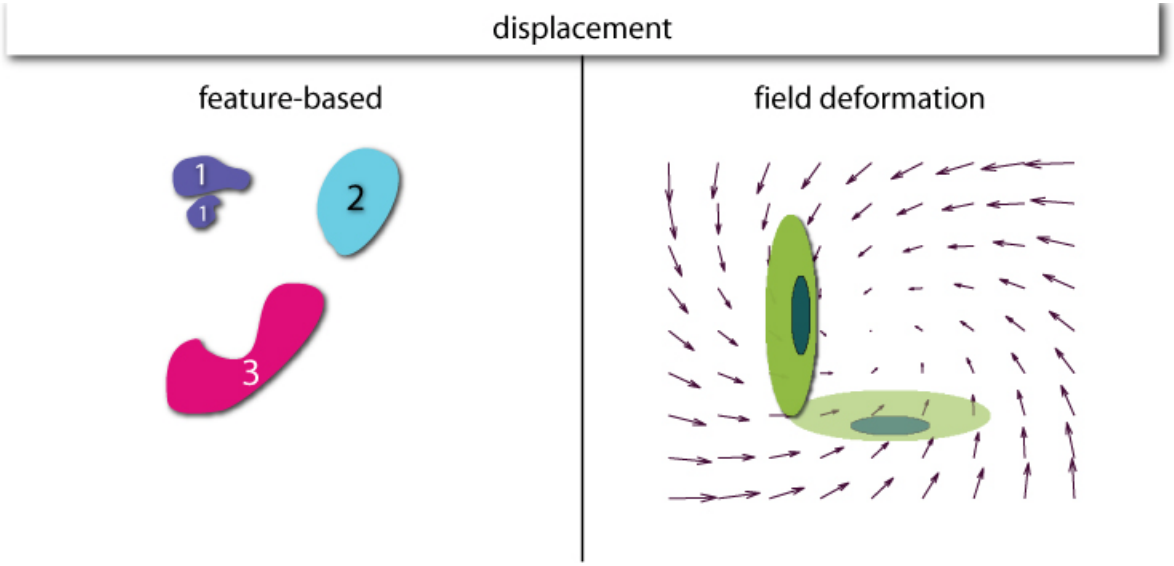
Limitations: Filtering methods

Does not clearly isolate specific errors (e.g., displacement, amplitude, structure)



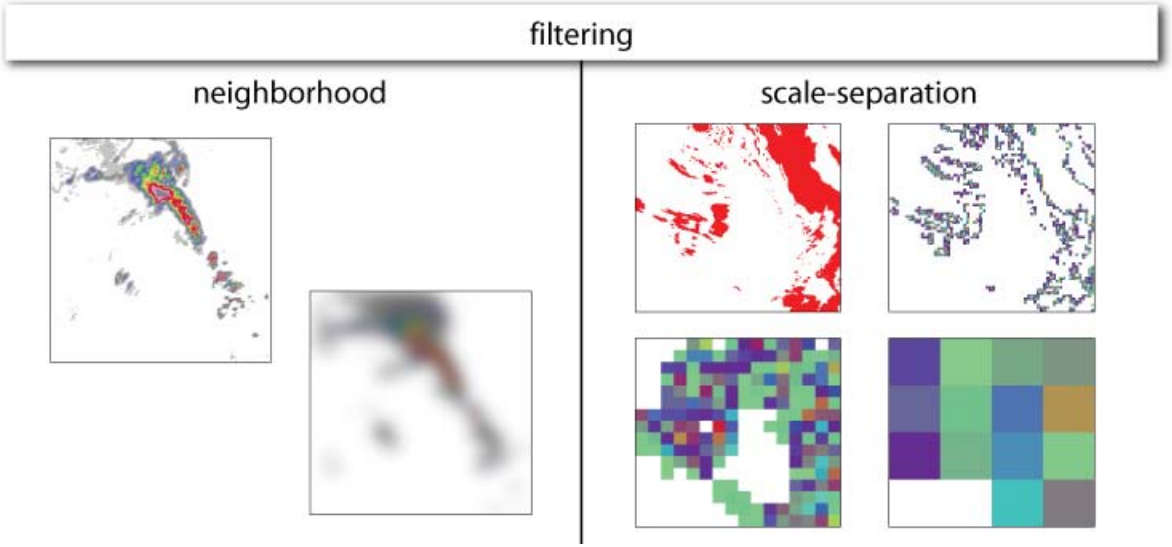
Limitations: Displacement methods

- May have somewhat arbitrary matching criteria
- Often many parameters to be defined
- More research needed on diagnosing mesoscale structure



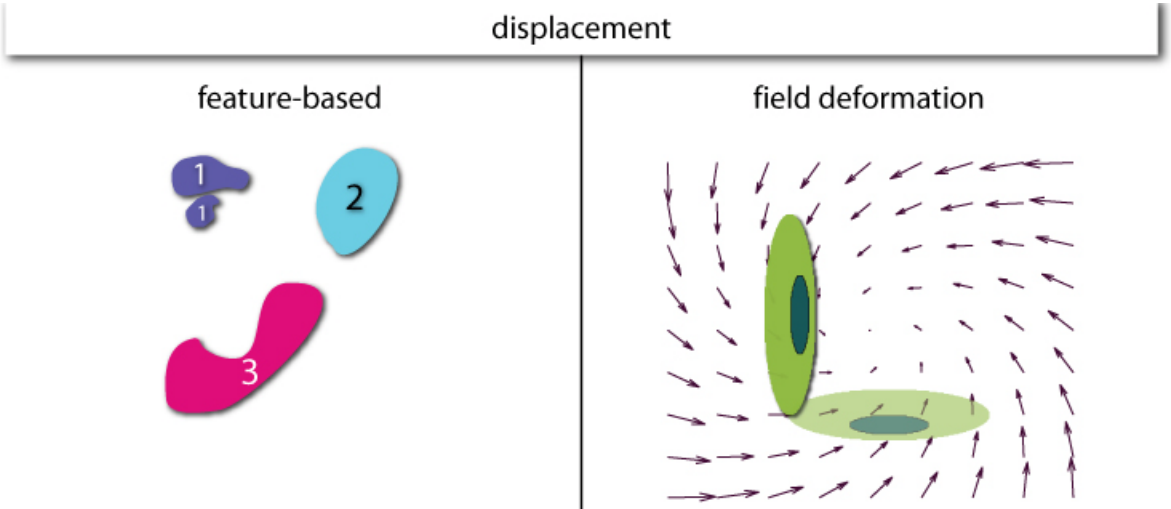
Strengths – Filtering methods

- Accounts for
  - Unpredictable scales
  - Uncertainty in observations
- Simple – ready-to-go
- Evaluates different aspects of a forecast (e.g., texture)
- Provides information about scale-dependent skill



Strengths – Displacement methods

- Features-based
  - Gives credit for close forecast
  - Measures displacement, structure
    - Provides diagnostic information
- Field-deformation
  - Can distinguish between aspect ratio and orientation angle error
  - Gives credit for a close forecast



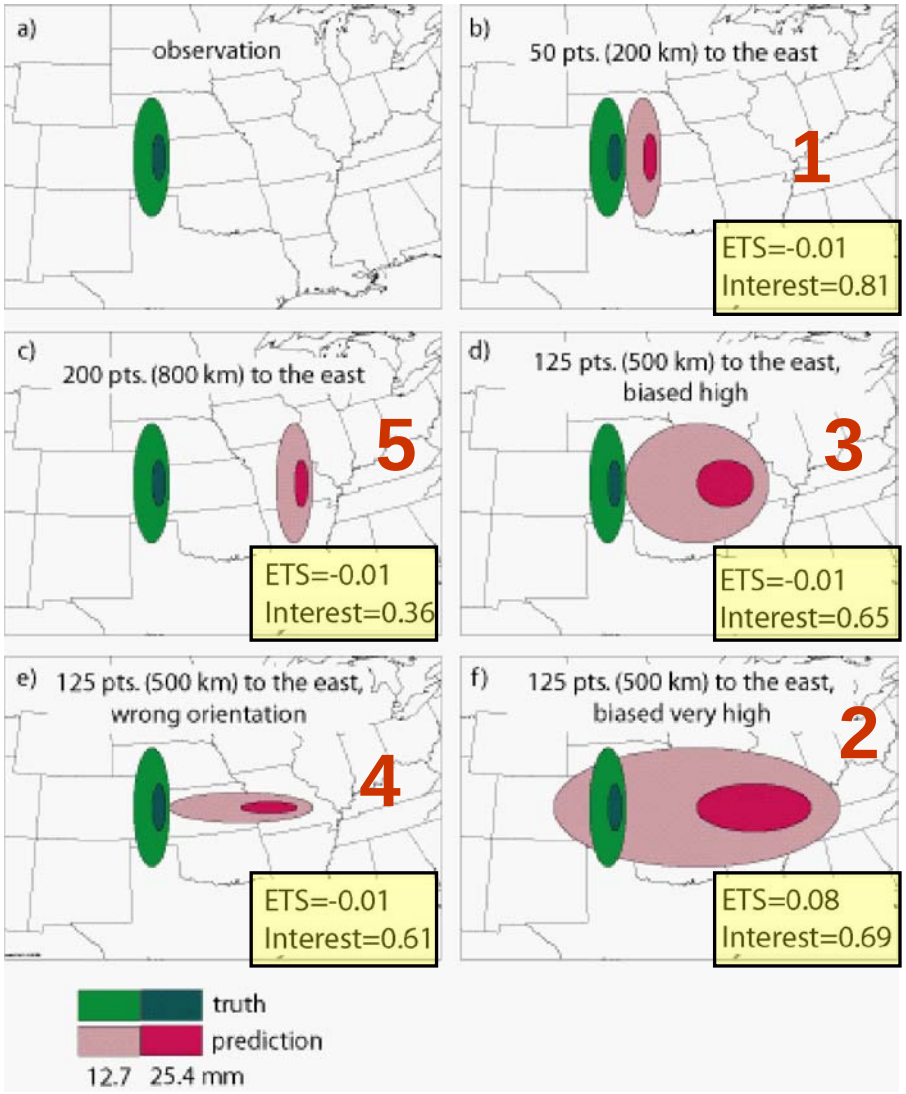
What do the new methods measure?

	Method	Scale-specific errors	Scales of useful skill	Structure errors	Location errors	Intensity errors	Hits, misses, false alarms, correct negatives
	Traditional	No	No	No	No	**Yes**	**Yes**
Filter	**Neighborhood**	**Yes**	**Yes**	No	Sensitive, but no direct information	**Yes**	**Yes**
Filter	**Scale Separation**	**Yes**	**Yes**	No	Sensitive, but no direct information	**Yes**	**Yes**
Displacement	**Field deformation**	No	No	No	**Yes**	**Yes**	No
Displacement	**Features-based**	Indirectly	No	**Yes**	**Yes**	**Yes**	Yes , based on features rather than gridpoints

Back to the original example… What can the new methods tell us?

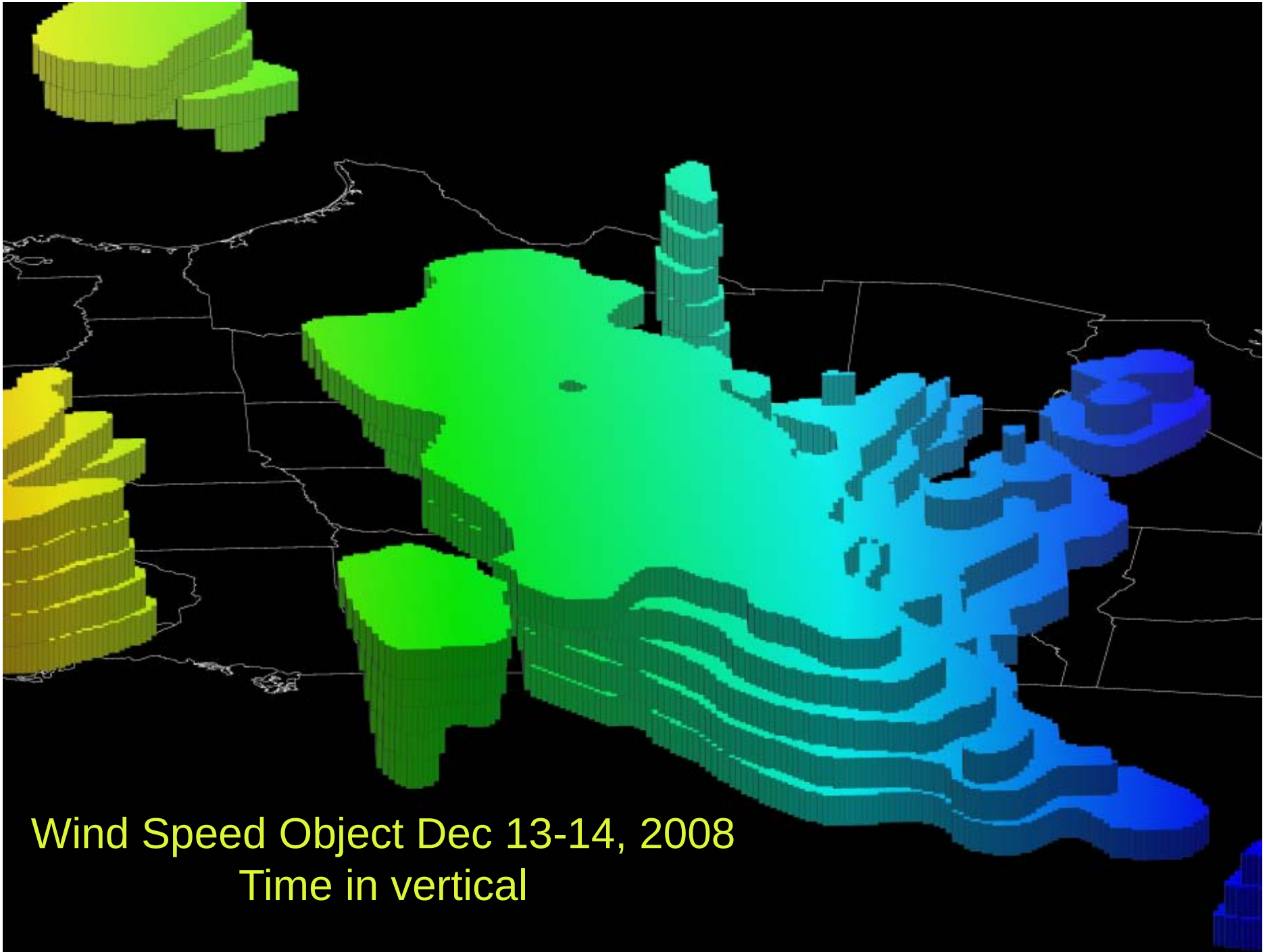
Example:

- MODE "Interest" measures overall ability of forecasts to match obs
- Interest values provide more intuitive estimates of performance than the traditional measure (ETS)
- But note: Even for spatial methods, Single measures don't tell the whole story!



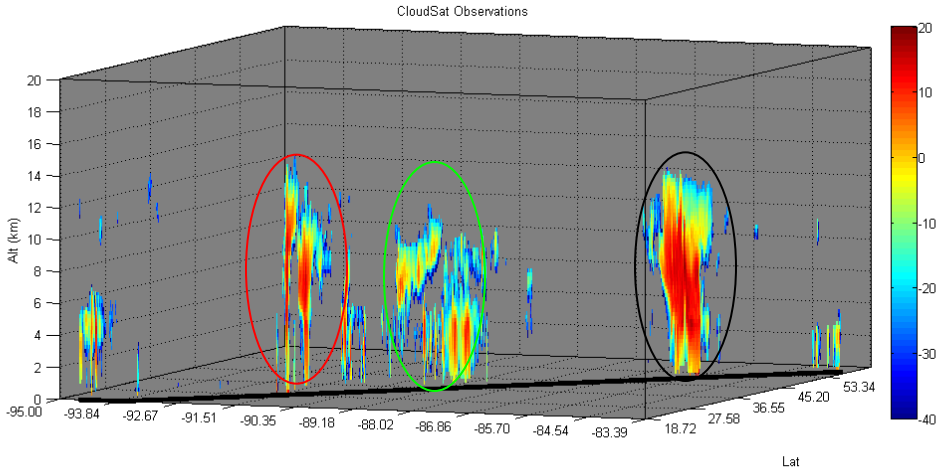
Application to other fields

- Methods have been commonly applied to precipitation and reflectivity
- New applications
  - Wind
  - Cloud analysis
  - Vertical cloud profile
  - Satellite estimates precipitation
  - Tropical cyclone structure

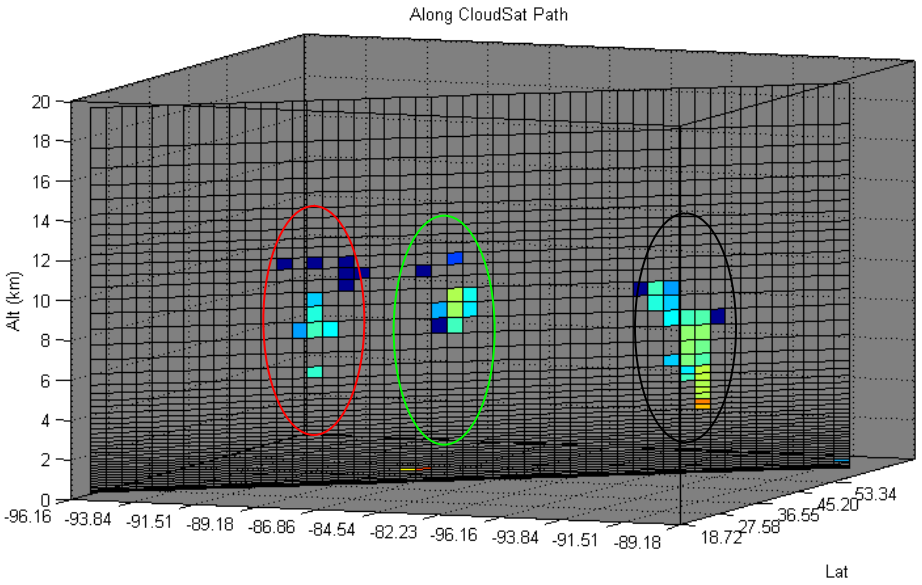

Wind Speed Object Dec 13-14, 2008
Time in vertical

Cloud-Sat Object-based Comparison: Along Track

CPR reflectivity

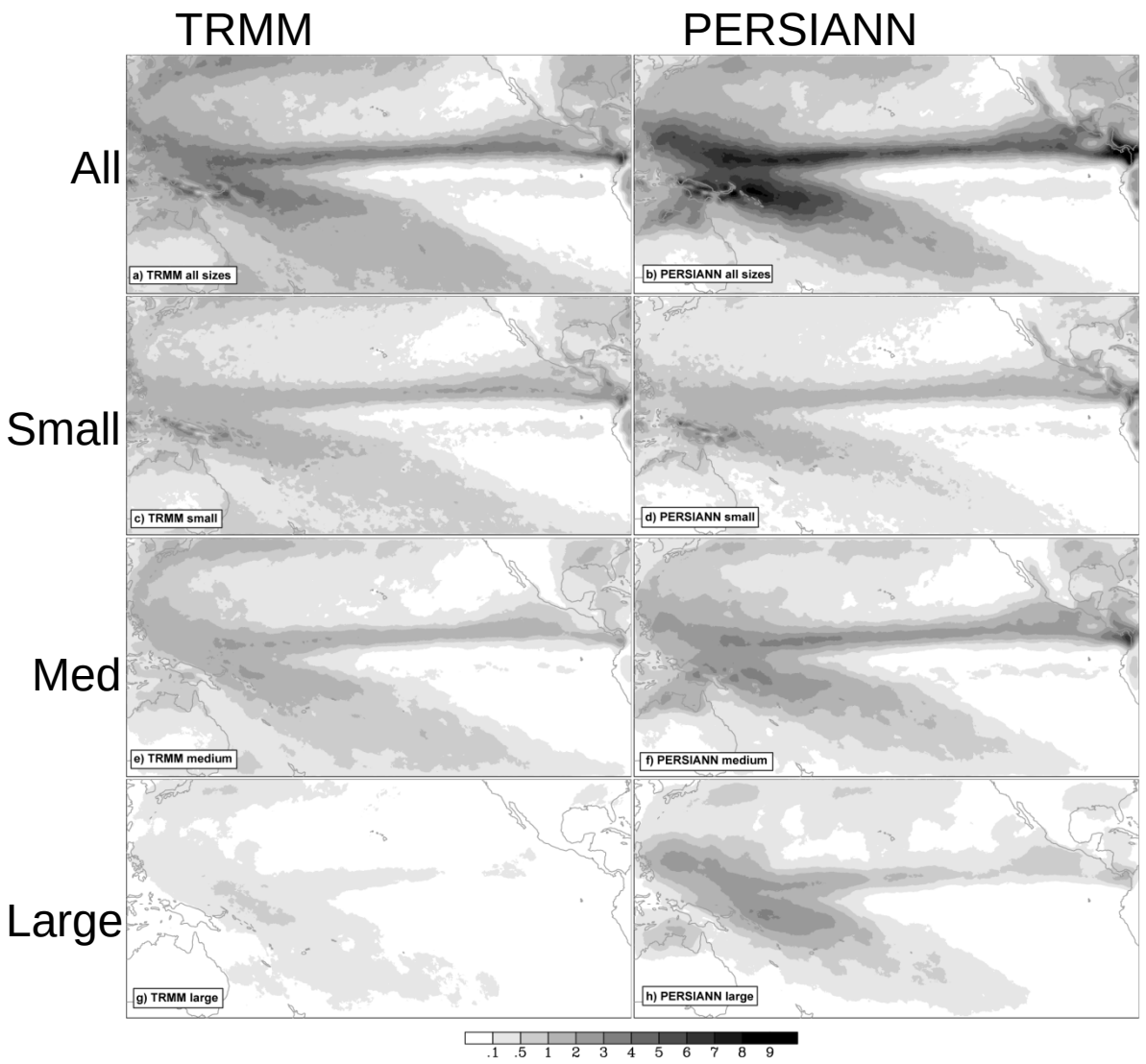


RUC reflectivity



Satellite precipitation estimates

Skok et al. (2010)
Object counts



Conclusion

- New spatial methods provide great opportunities for more meaningful evaluation of spatial fields
  - Feed back into forecast or product development
  - Measure aspects of importance to users
- Each method is useful for particular types of situations and for answering particular types of questions
- Methods are useful for other types of fields
- For more information (and references), see http://www.rap.ucar.edu/projects/icp/index.html

Method availability

- Neighborhood, Intensity-Scale, and MODE methods are available as part of the Model Evaluation Tools (MET)
  - Available at http://www.dtcenter.org/met/users/
  - Implemented and supported by the Developmental Testbed Center and staff at the NCAR/RAL/JNT
- Software for other methods may be available on the ICP web page

  http://www.ral.ucar.edu/projects/icp/index.html

  or directly from the developer