Spatial Forecast Verification Methods

Barbara Brown

Joint Numerical Testbed Program

Research Applications Laboratory, NCAR

22 October 2014

<u>Acknowledgements</u>: Tara Jensen, Randy Bullock, Eric Gilleland, David Ahijevych, Beth Ebert, Marion Mittermaier, Mike Baldwin

Goal

Describe new methods for evaluation of spatial fields

Many methods have been developed in the context of high resolution precipitation forecasts, which may have application to aerosol predictions

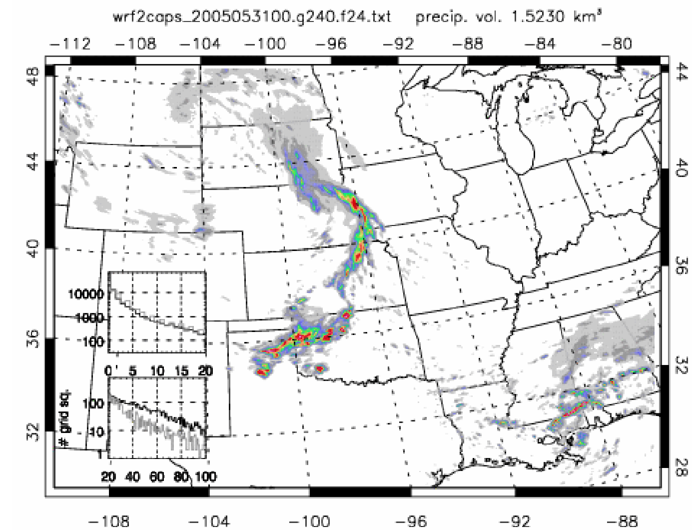
Context

- Motivation: Traditional verification approaches don't
  - Reflect observed capabilities of new (higher-resolution) modeling systems (generally with higher spatial resolution)
  - Provide diagnostic information about model performance (i.e., *what* was wrong)
- History: Most new spatial methods developed over the last 10-15 years
  - Initial methods published in early 1990s
  - Earliest "used" method published in 2000 (Ebert and McBride)
  - Beginning to be used operationally
- Initial development target: Mesoscale precipitation
- Other applications:
  - Clouds and Reflectivity
  - Jets and Low pressure systems (Mittermaier)
  - Convective storm characteristics (Clark)
  - Wind/RH
  - Climate/climatology
- Limitations: Typically requires gridded forecasts and observations

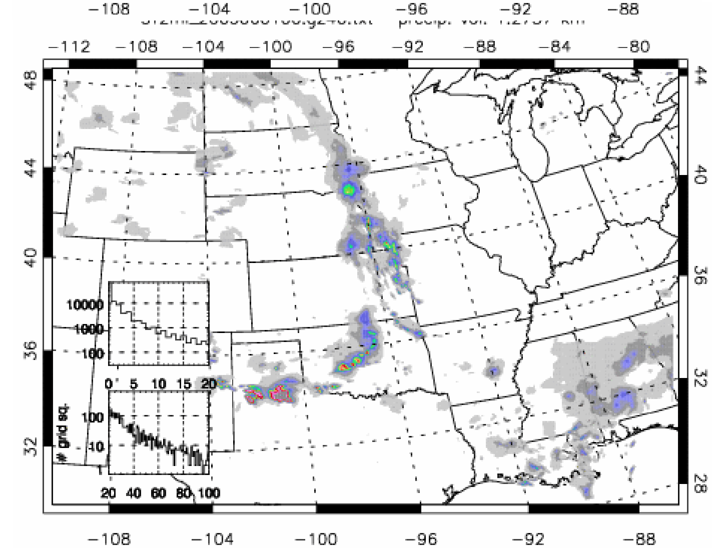
Spatial fields

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

WRF model

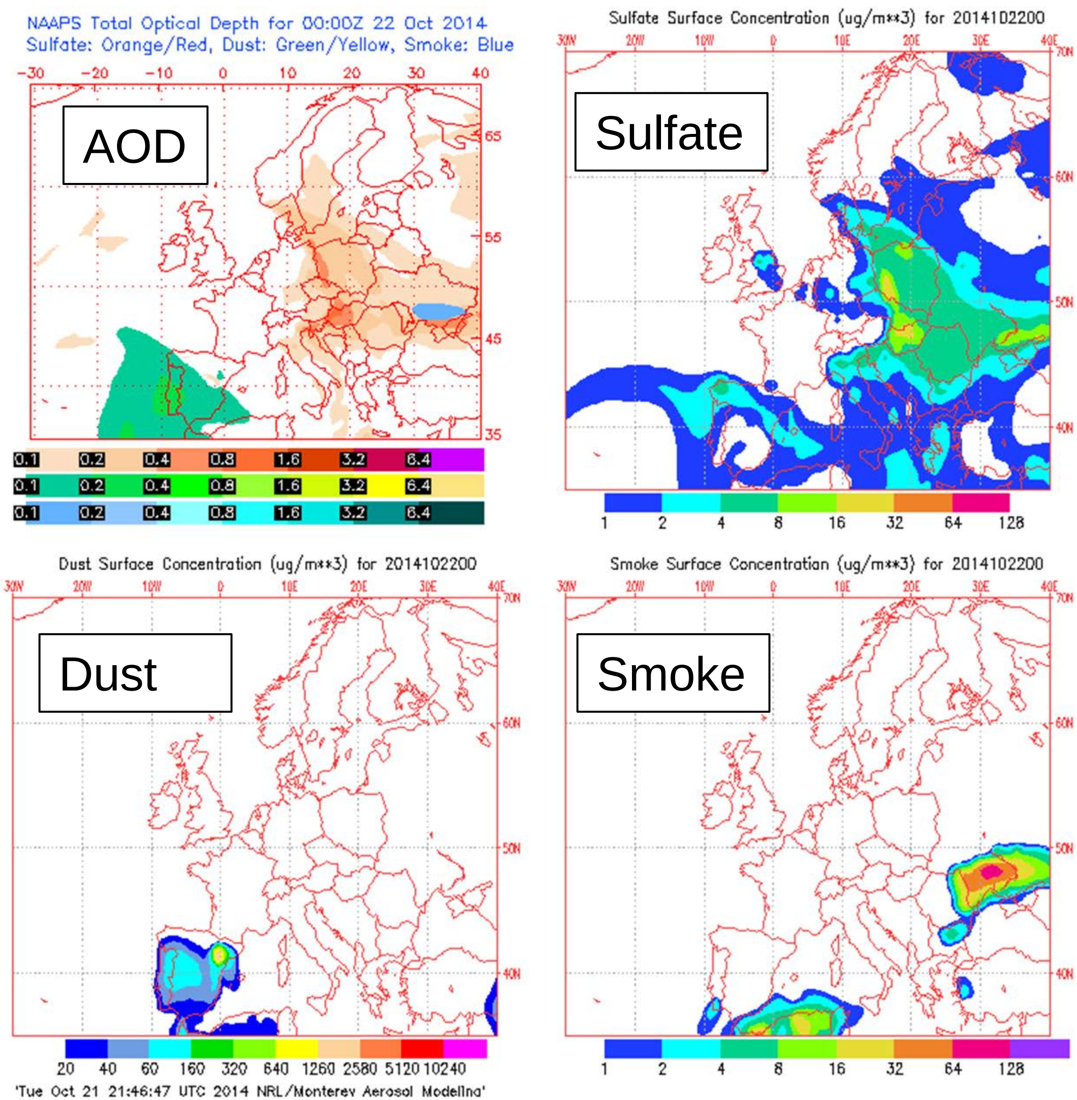


Stage II radar



Spatial fields

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

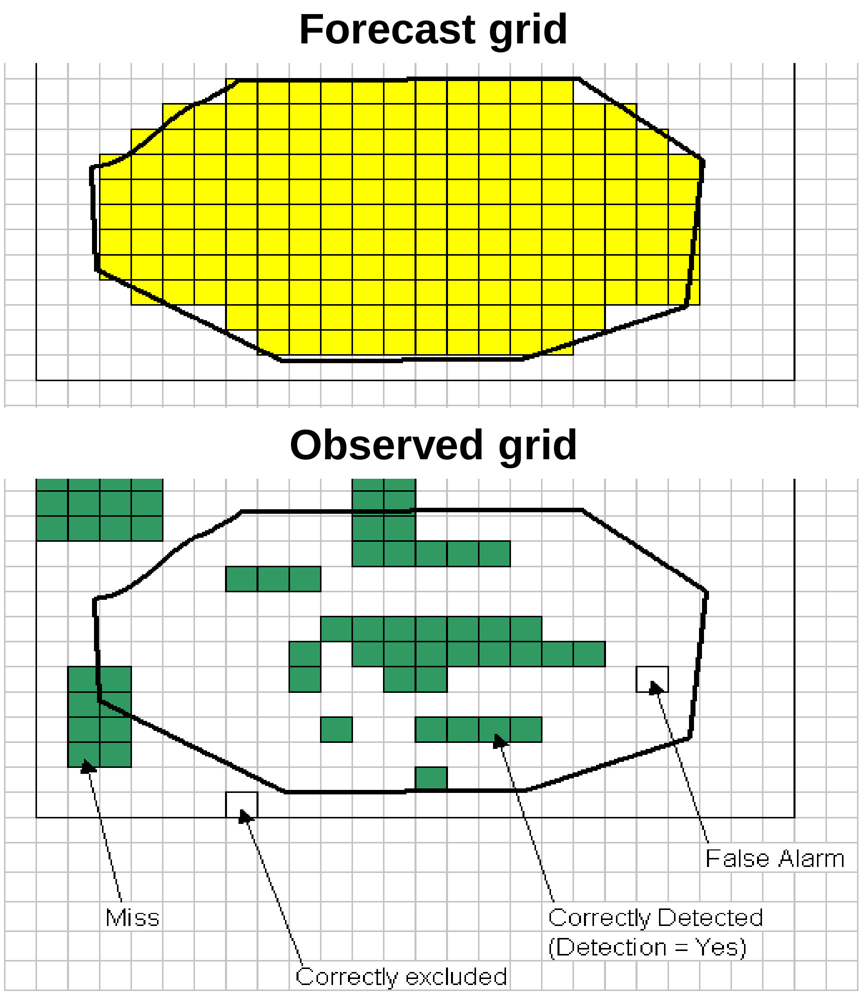


Matching two fields (forecasts and observations)

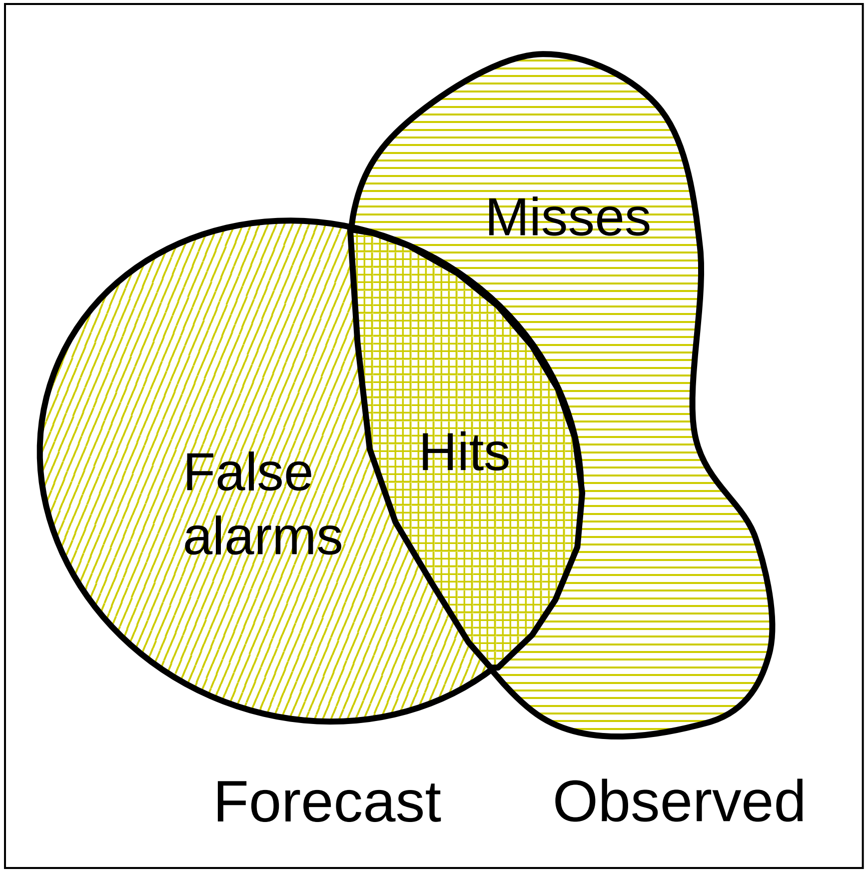
<u>Focus</u>: Gridded fields

Traditional grid to grid approach:

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



Traditional spatial verification measures

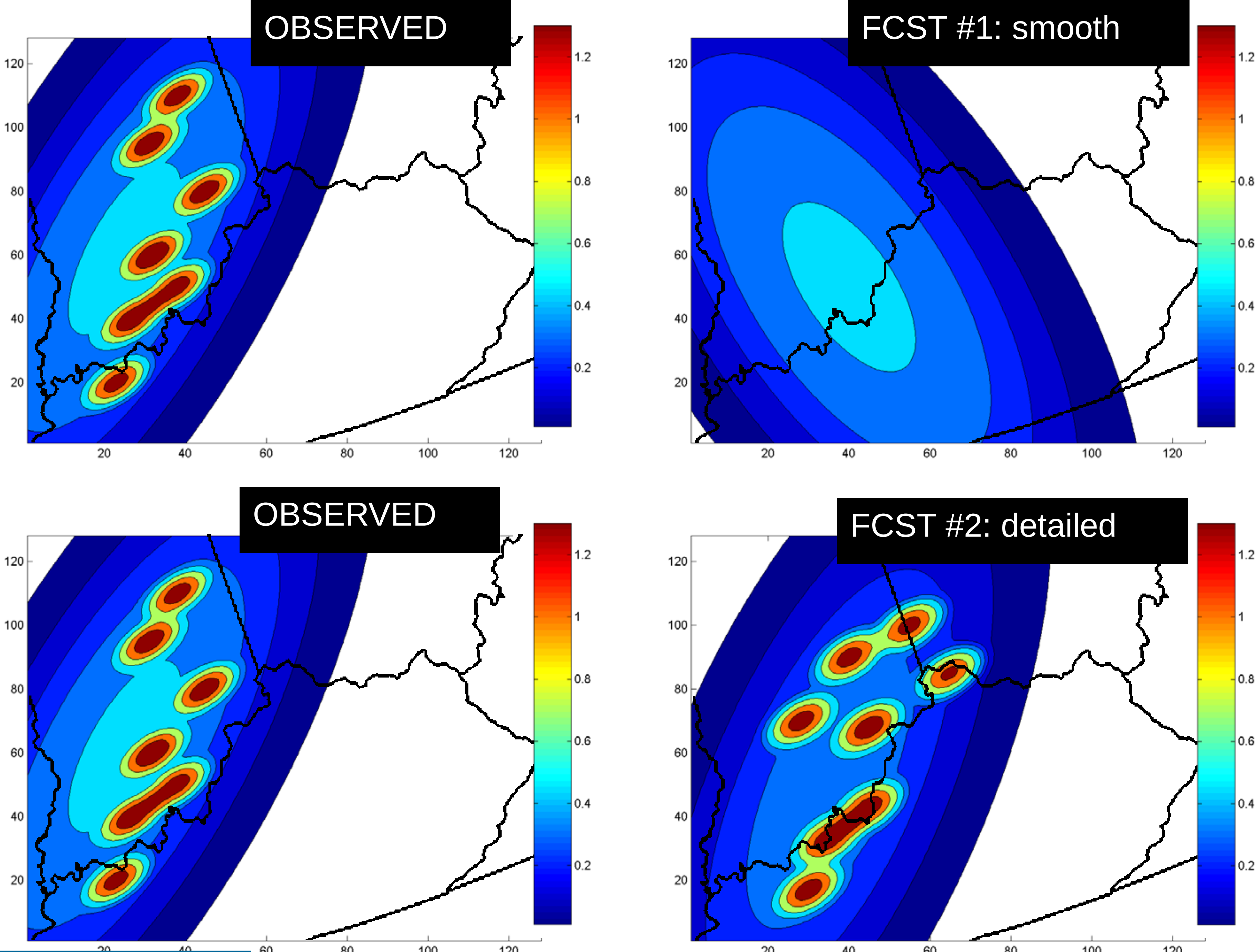


Contingency Table

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

Basic methods:

1. Create contingency table by thresholding forecast and observed values
   - *Compute traditional contingency table statistics: POD, FAR, Freq. Bias, CSI, GSS (= ETS)*
2. Directly compute errors in predictions
   - *Compute measures for continuous variables*: MSE, MAE, ME



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

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?

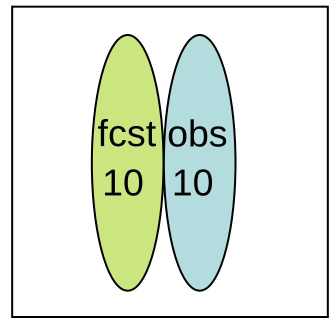
Double penalty problem

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

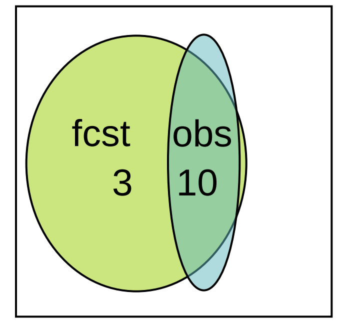
<u>Double penalty:</u>

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

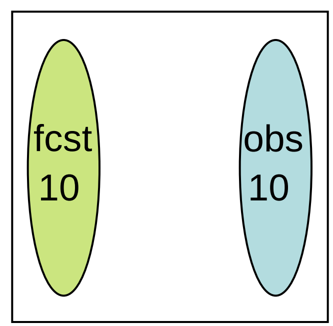
(2) No event predicted where it did occur => ***Miss***



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



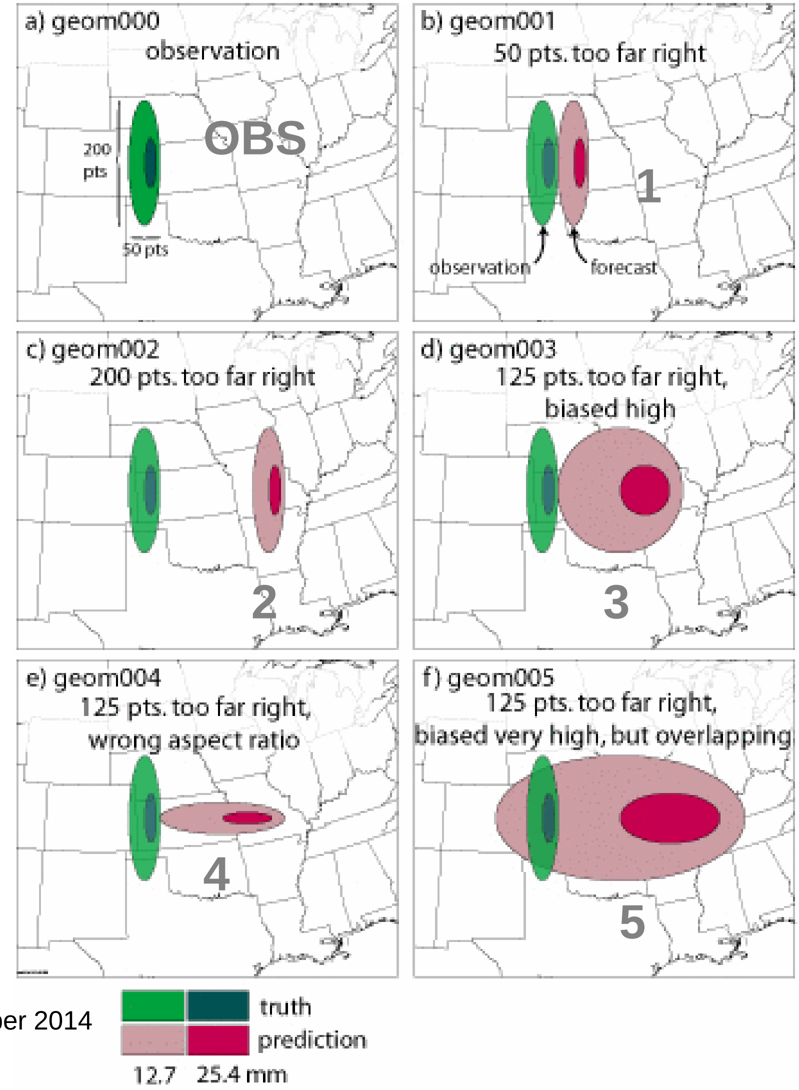
Summary: What are the issues with the traditional approaches?

- “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

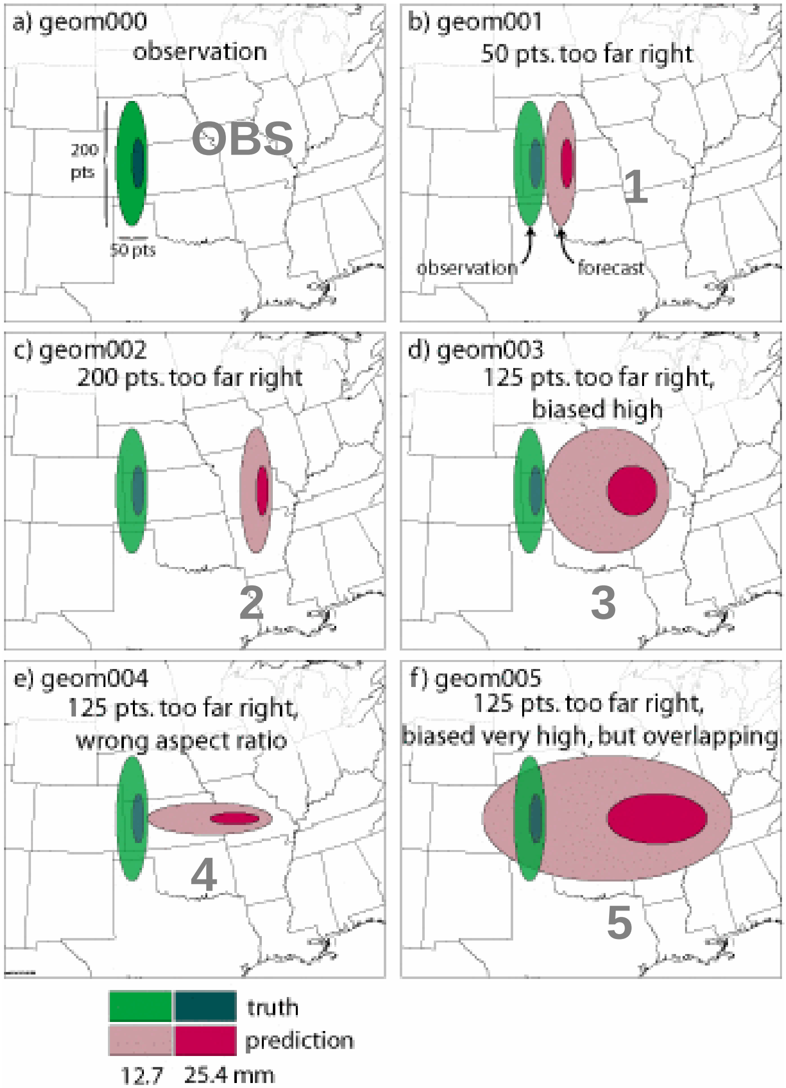
Scores for Examples 1-4:

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"

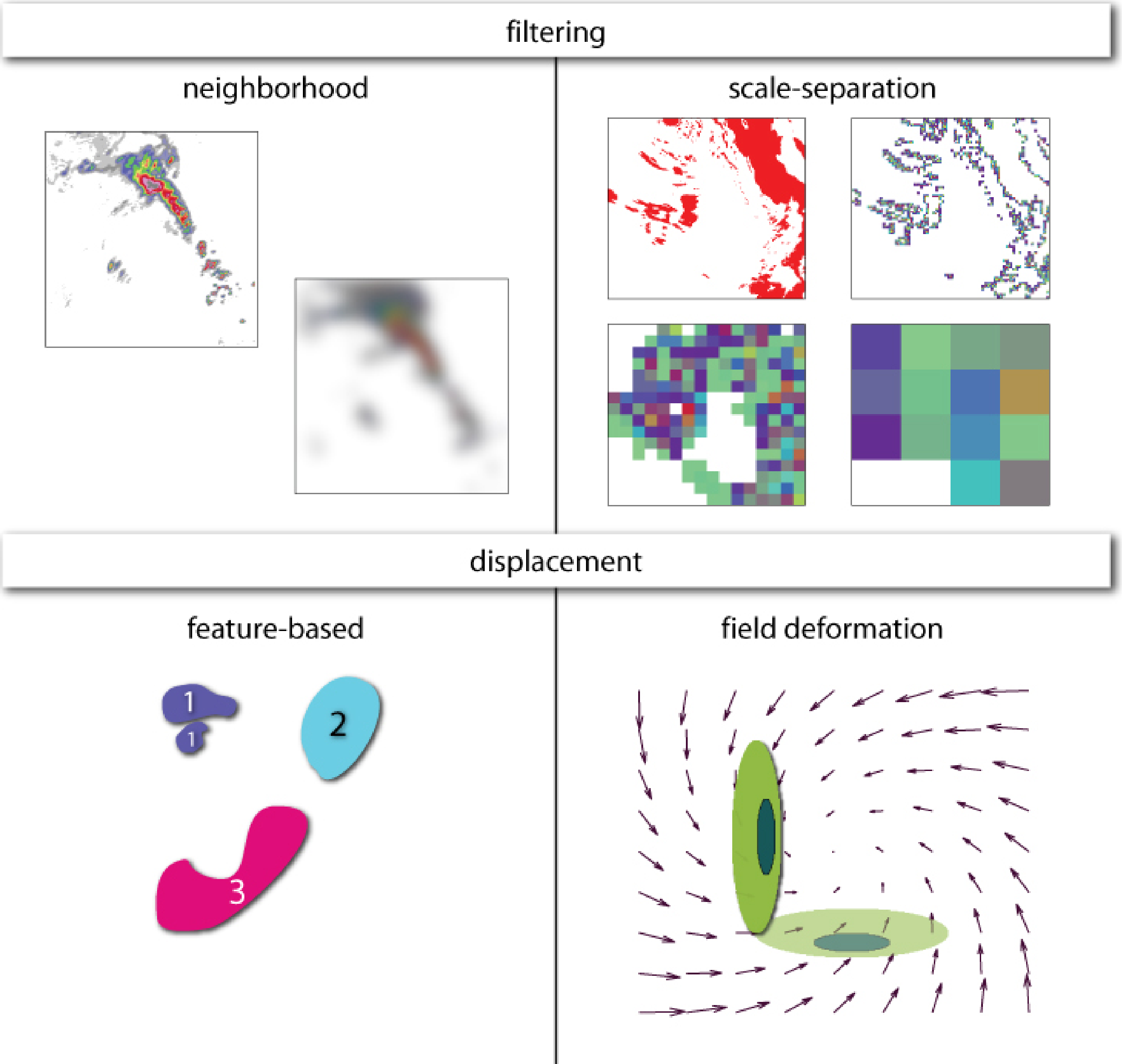


Summary: What are the issues with the traditional approaches?

- "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?

Spatial Method Categories

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



New spatial verification approaches

Neighborhood

Successive smoothing of forecasts/obs

Gives credit to "close" forecasts

Scale separation

Measure scale-dependent error

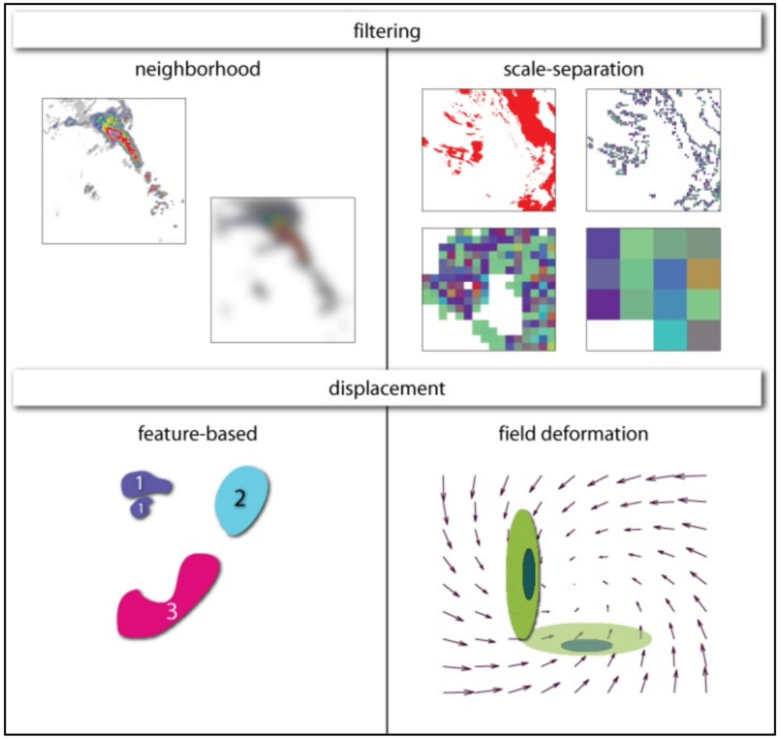
Field deformation

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

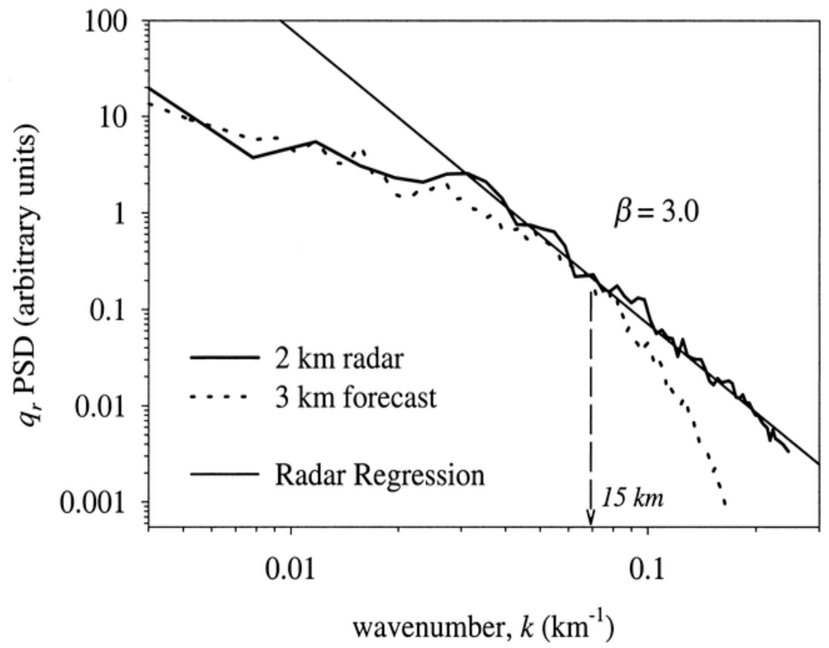
Evaluate attributes of identifiable features



Scale separation methods

- Goal:
  Examine performance as a function of spatial scale

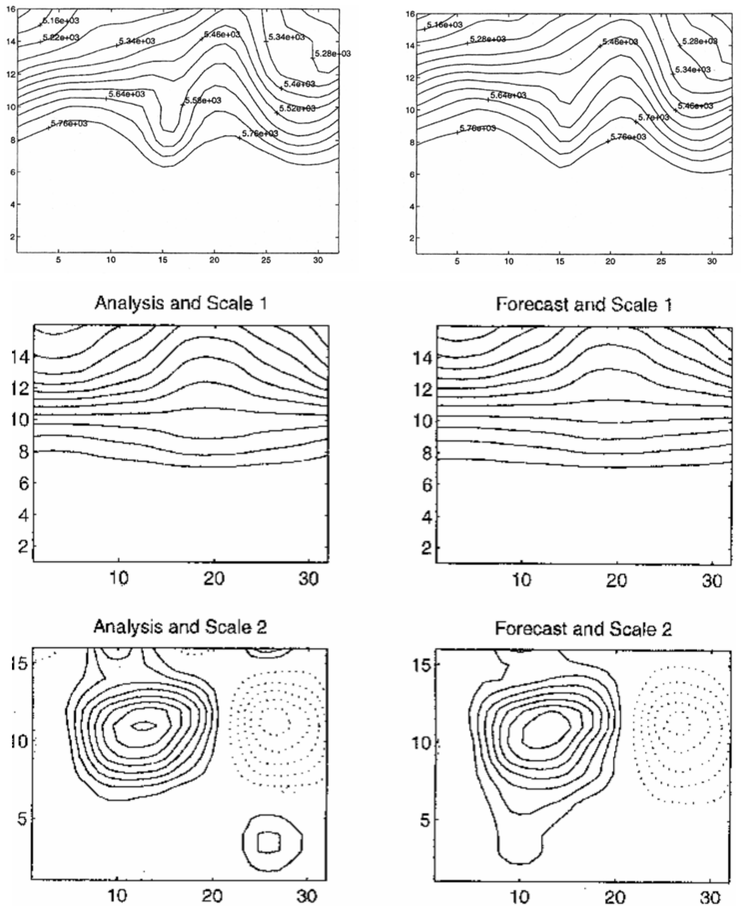
- Example: Power spectra
  - Does it look real?
  - Harris et al. (2001): compare multi-scale statistics for model and radar data



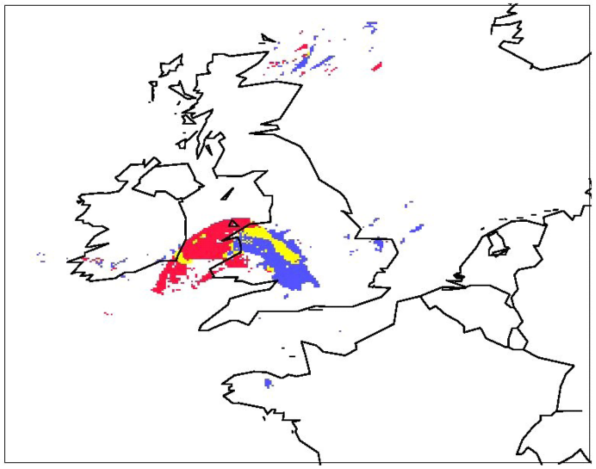
From Harris et al. 2001

Scale decomposition

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

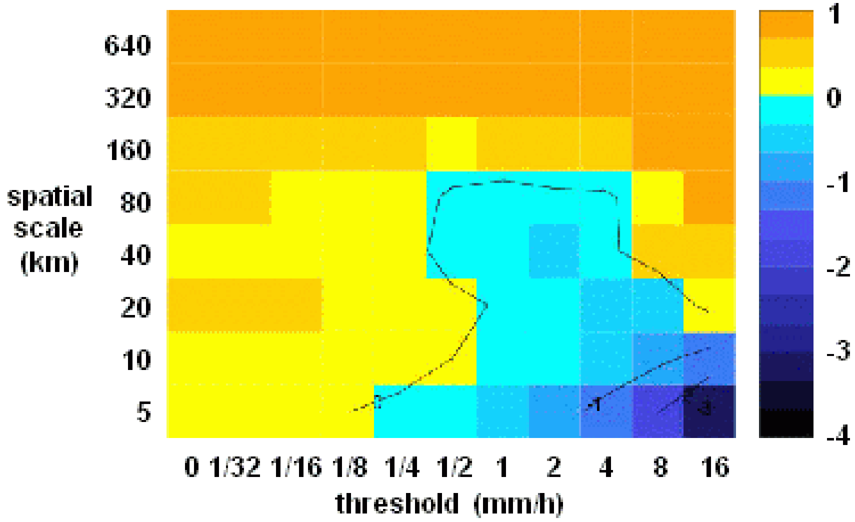


Scale separation methods



- 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)

MSE Skill Score

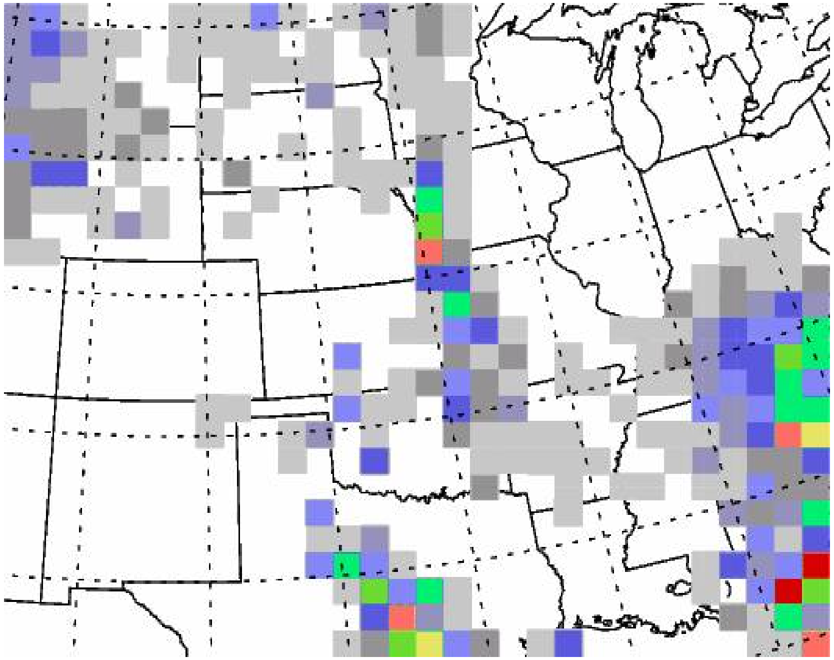


Neighborhood verification

<u>Goal:</u>

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

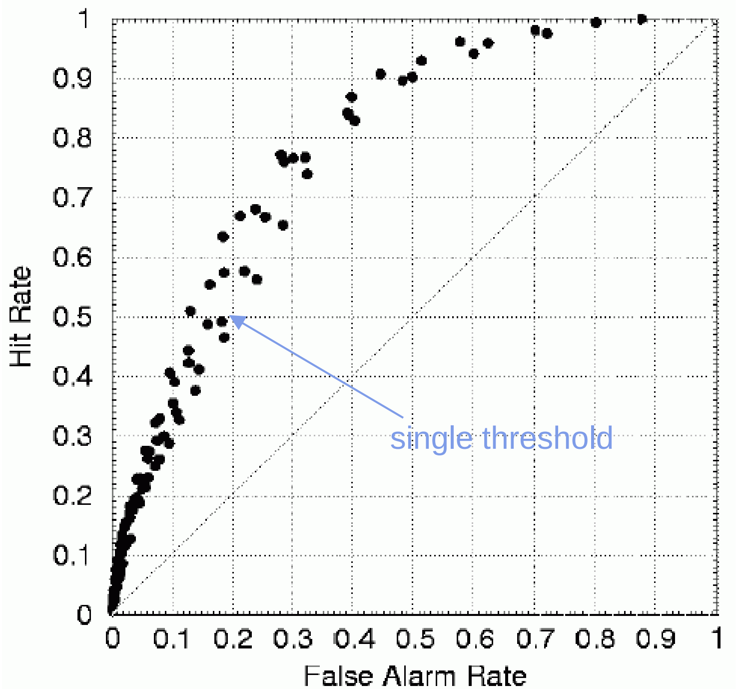
- Also called "fuzzy" verification
- Example: Upscaling
  - Put observations and/or forecast on coarser grid
  - Calculate traditional metrics
- Provide information about scales where the forecasts have skill



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)

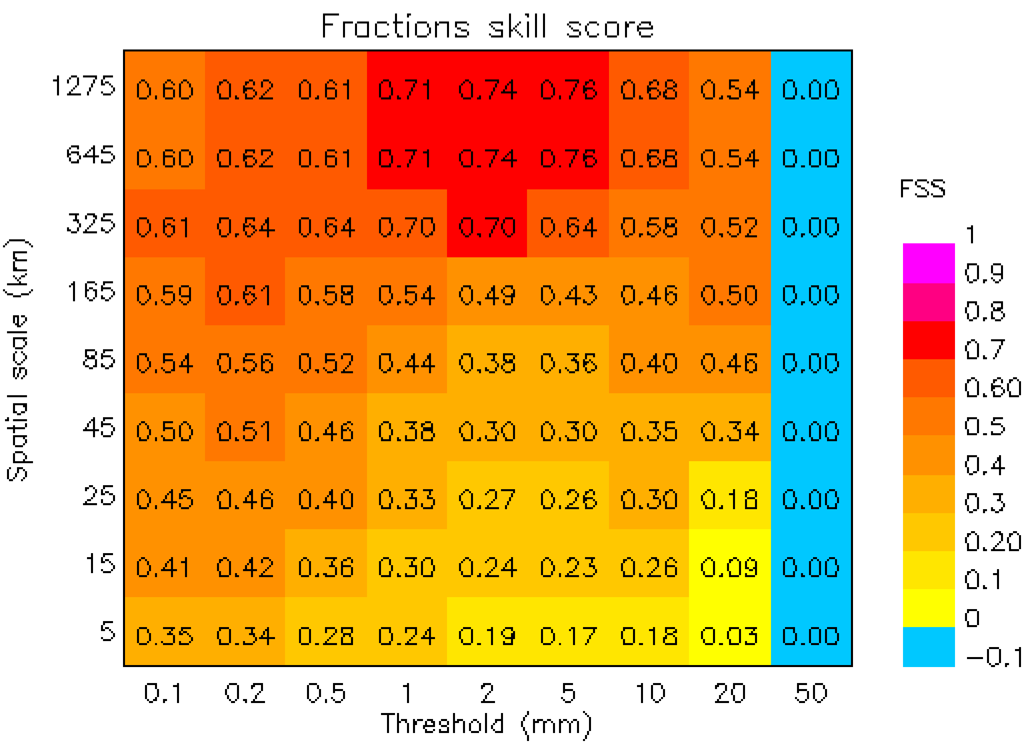


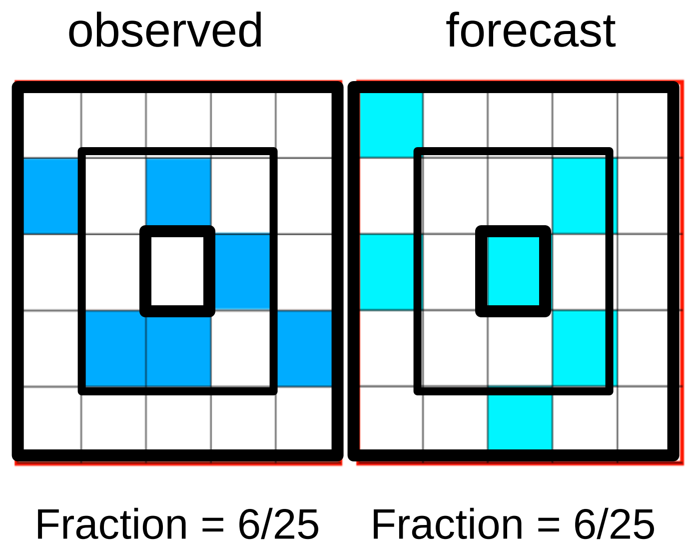
Atger, 2001

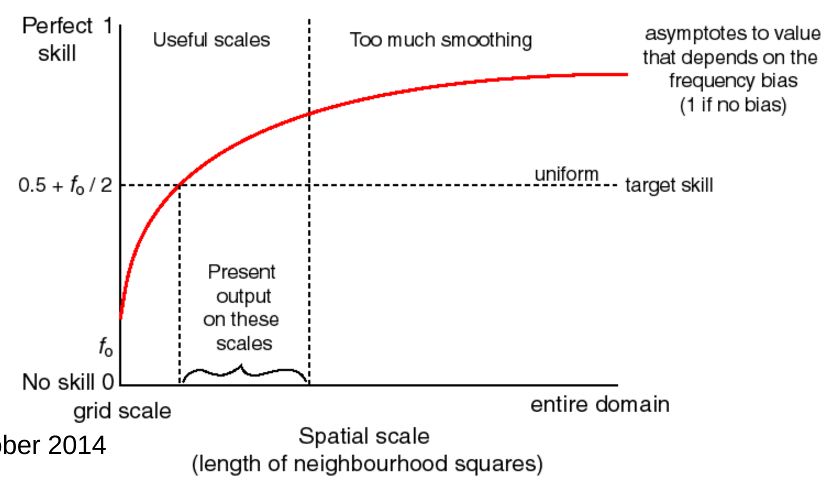
Ebert (2007; Met Applications) provides a review and synthesis of these approaches

Fractions skill score

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





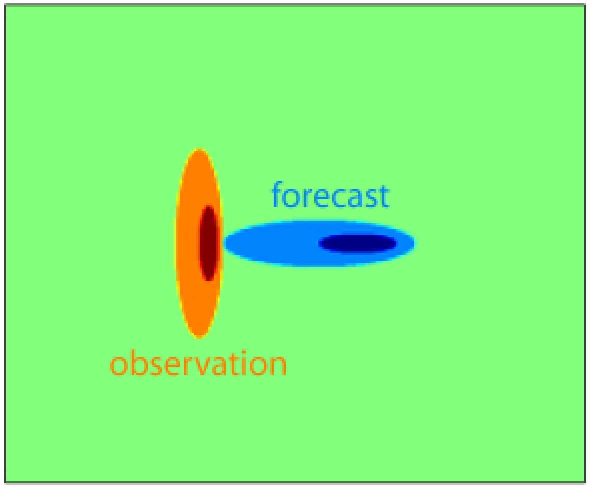


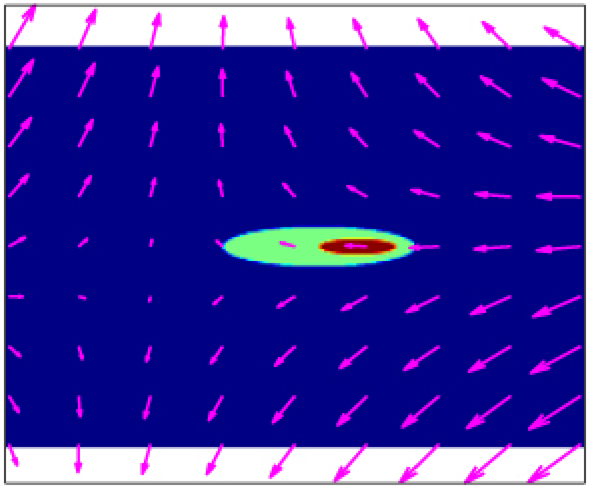
(Roberts 2005; Roberts and Lean 2007)

Field deformation

Goal:

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

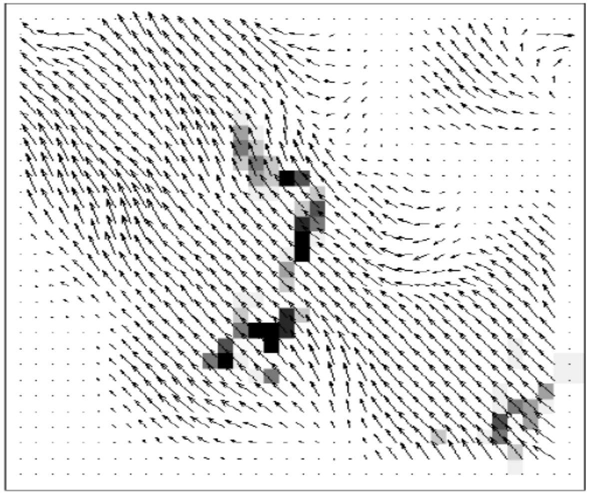




Field deformation methods

<u>Example methods :</u>

- 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)

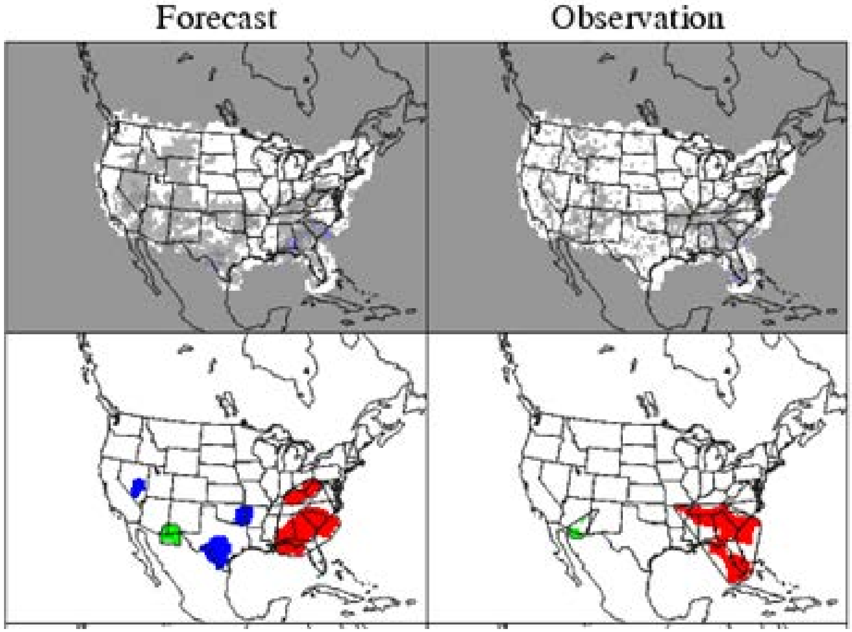


From Keil and Craig 2008

Object/Feature-based

Goals:

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



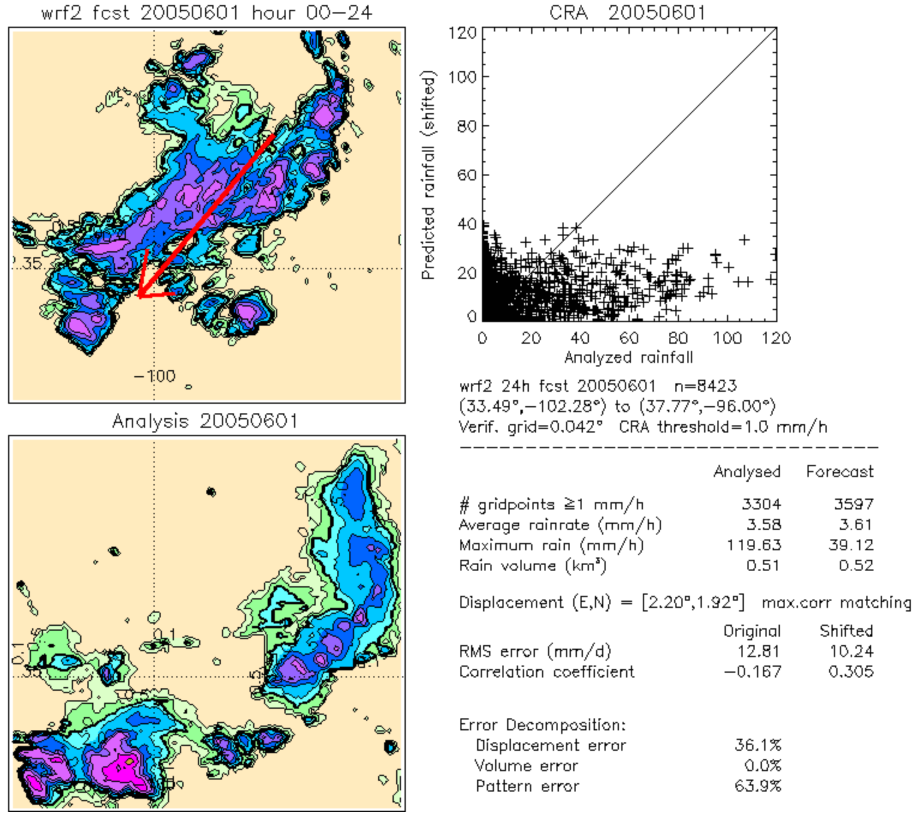
MODE example

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)

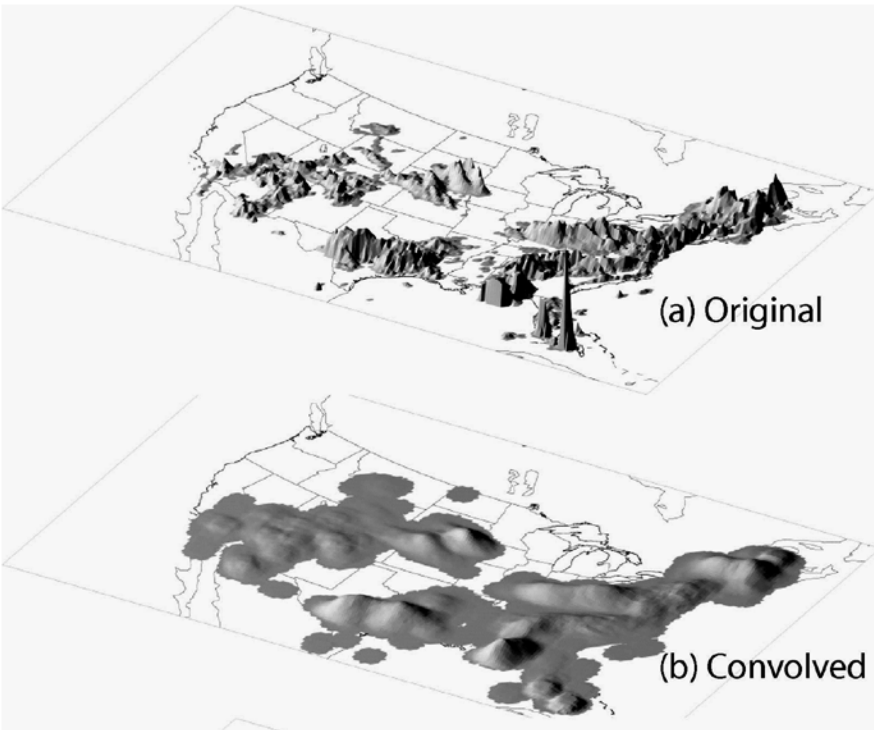
CRA example (Ebert and Gallus)

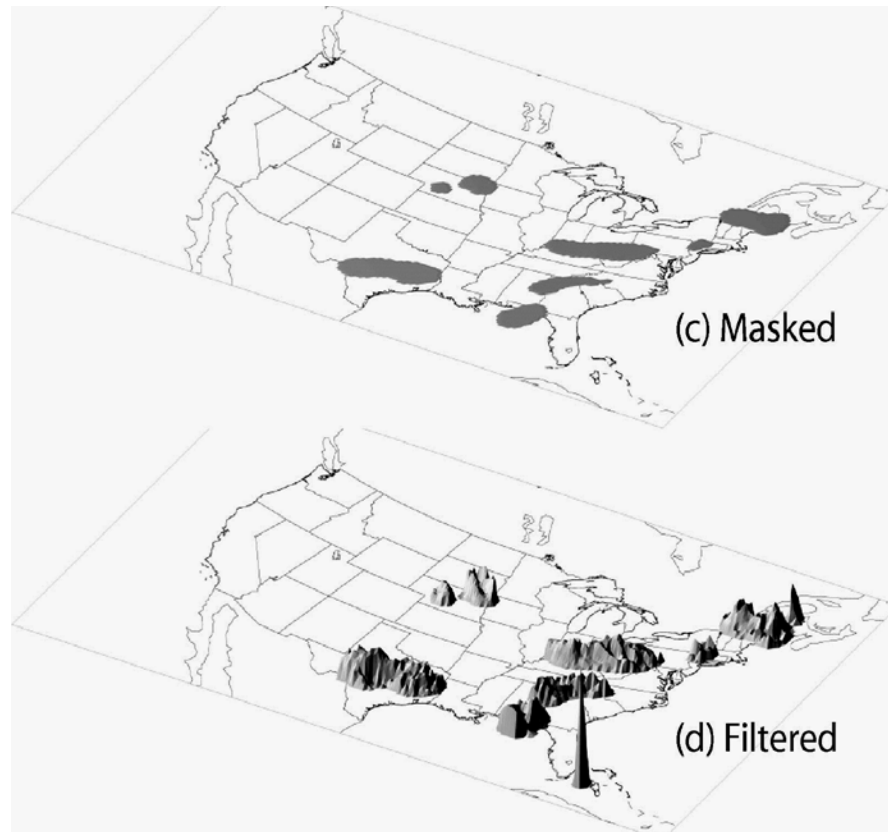


The CRA method measures displacement and estimates error due to displacement, pattern, and volume

MODE – Method for Object-based Diagnostic Evaluation

MODE Object identification



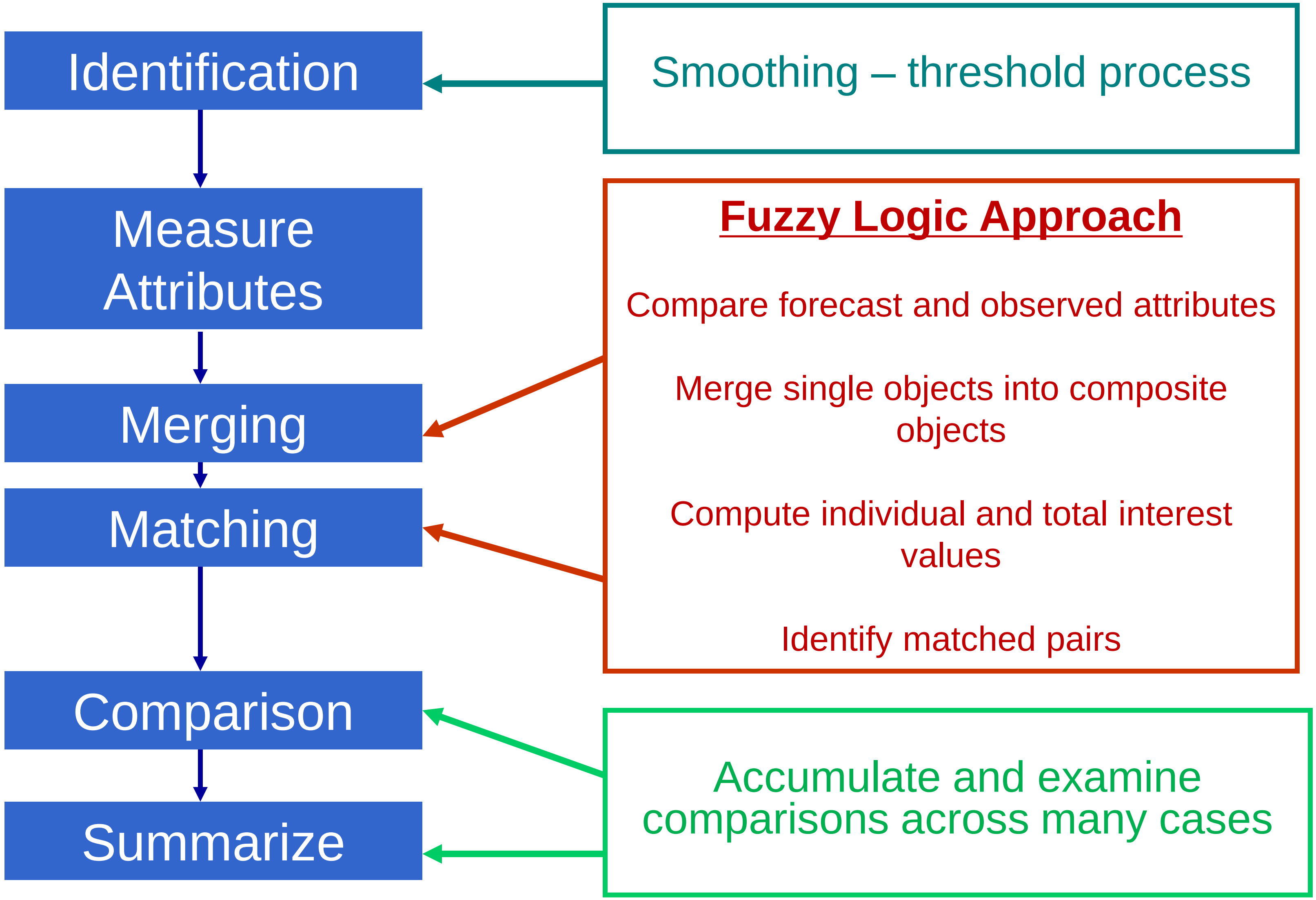


Two parameters:

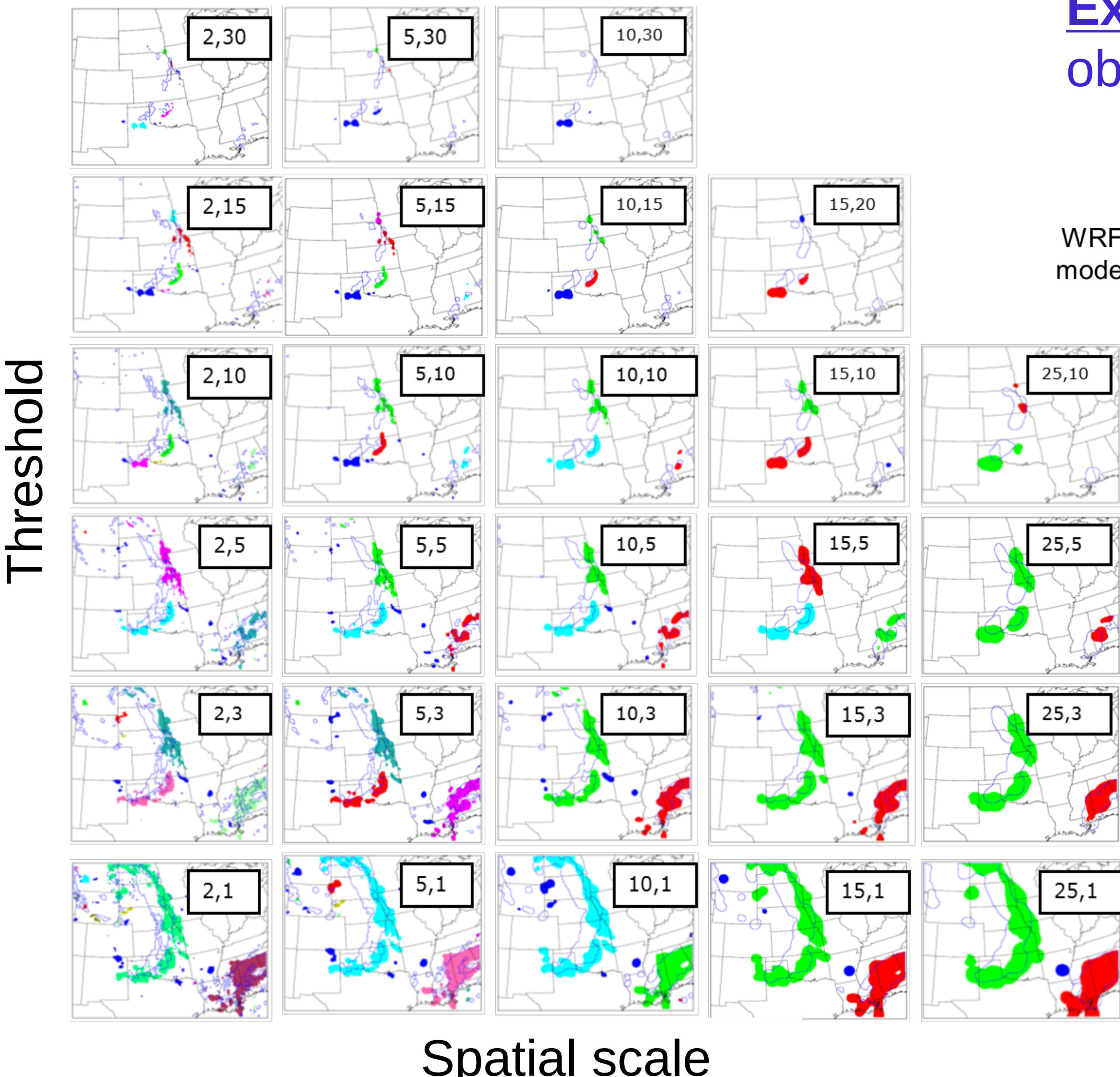
1. Convolution radius
2. Threshold

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

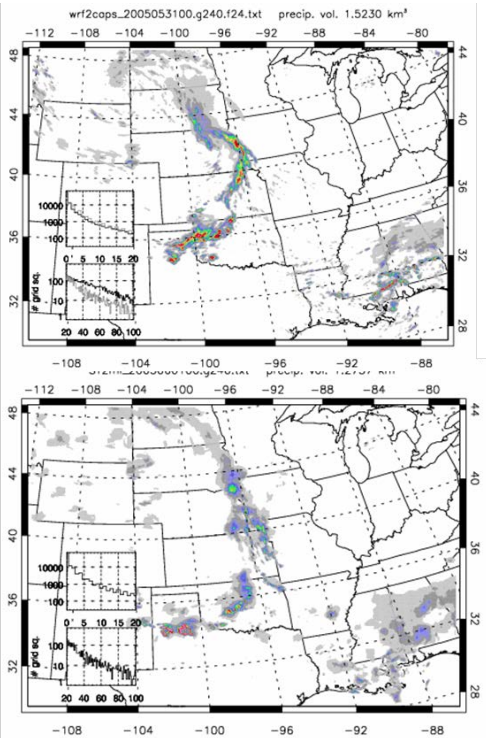
MODE methodology



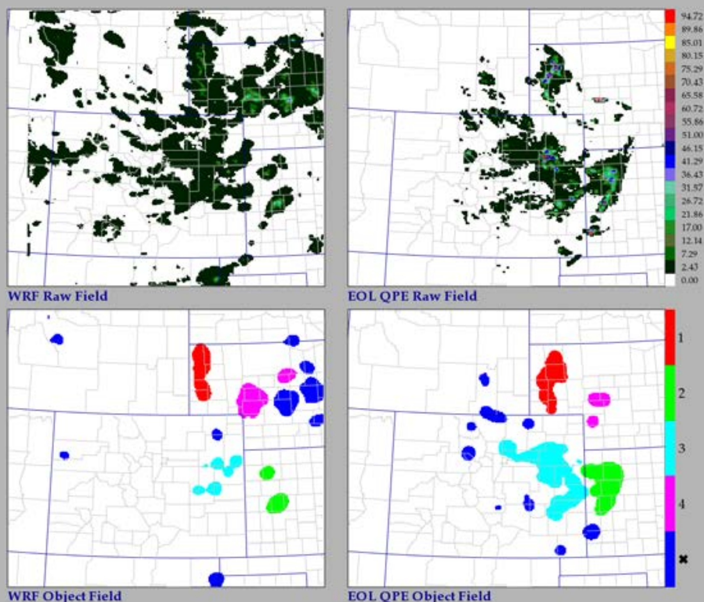
Example: MODE object definitions



WRF model



MODE Example: Summer 2014 QPF experiment (STEP)

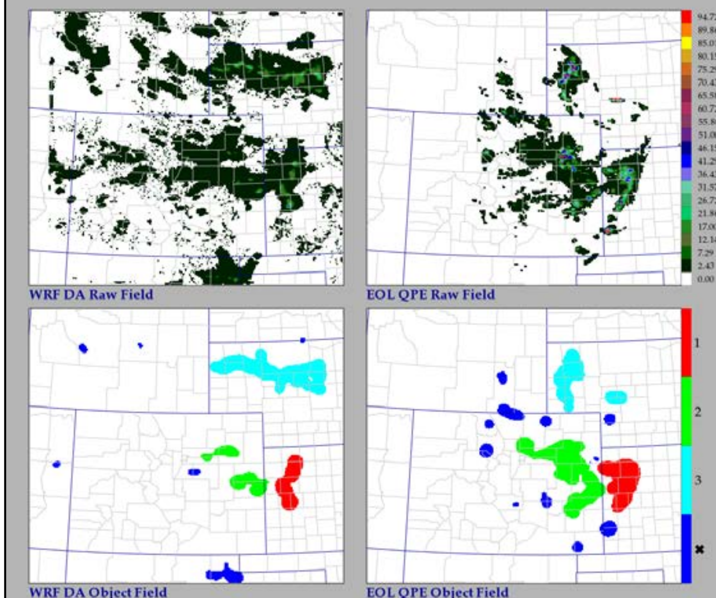


WRF vs. EOL QPE

Valid August 10, 2014 00h — **Radius** 6
Lead 03h — **Threshold** >=1.000

	Simple Objects	Cluster Objects	Simple Area	Cluster Area	Unmatched Area	10 %	50 %	90 %
Fcst	18	4	5,021	3,094	1,927	0.06	1.65	10.18
Obs	16	4	7,079	5,833	1,246	0.00	1.31	21.22

Cluster	Area (F/O)	Centroid Distance	Overlap Area	Symmetric Difference	50 % (F/O)	90 % (F/O)	Total Interest
	858 / 1,149	12.76	424	1,583	2.01 / 2.83	7.68 / 34.91	0.99
	536 / 1,477	13.82	267	1,746	1.75 / 4.22	14.18 / 26.82	0.93
	666 / 2,813	8.39	327	3,152	1.21 / 2.17	5.32 / 18.35	0.85
	1,034 / 394	11.74	263	1,165	2.20 / 17.70	14.13 / 78.17	0.93



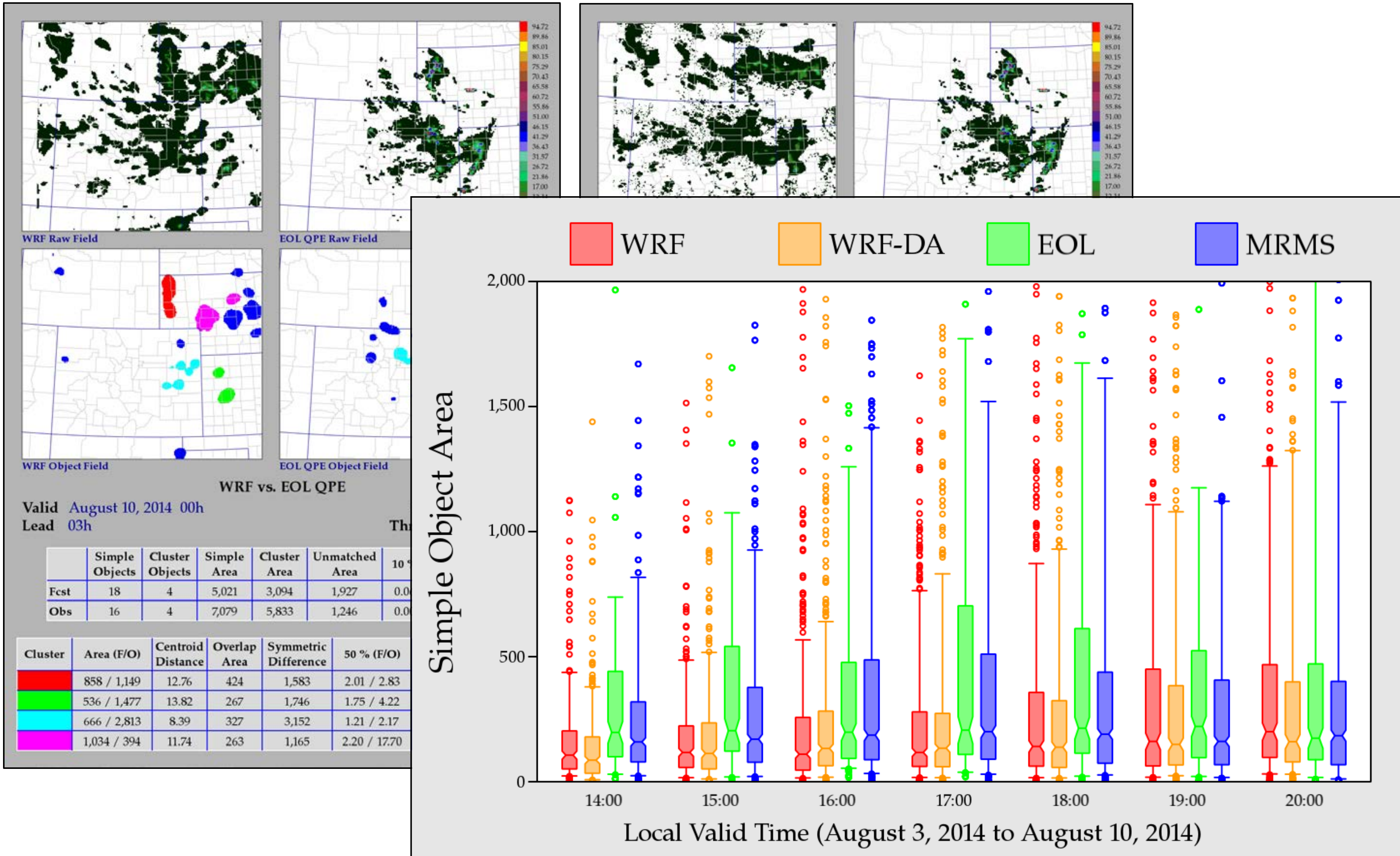
WRF DA vs. EOL QPE

Valid August 10, 2014 00h — **Radius** 6
Lead 03h — **Threshold** >=1.000

	Simple Objects	Cluster Objects	Simple Area	Cluster Area	Unmatched Area	10 %	50 %	90 %
Fcst	10	3	5,042	4,404	638	0.09	1.58	10.00
Obs	16	3	7,079	5,698	1,381	0.00	1.31	21.22

Cluster	Area (F/O)	Centroid Distance	Overlap Area	Symmetric Difference	50 % (F/O)	90 % (F/O)	Total Interest
	909 / 1,477	6.74	738	1,648	1.42 / 4.22	12.24 / 26.82	0.97
	919 / 2,794	9.58	777	2,936	1.35 / 2.18	4.66 / 18.37	0.93
	2,576 / 1,427	38.01	406	3,597	1.90 / 3.66	11.60 / 40.19	0.95

MODE Example: Summer 2014 QPF experiment (STEP)



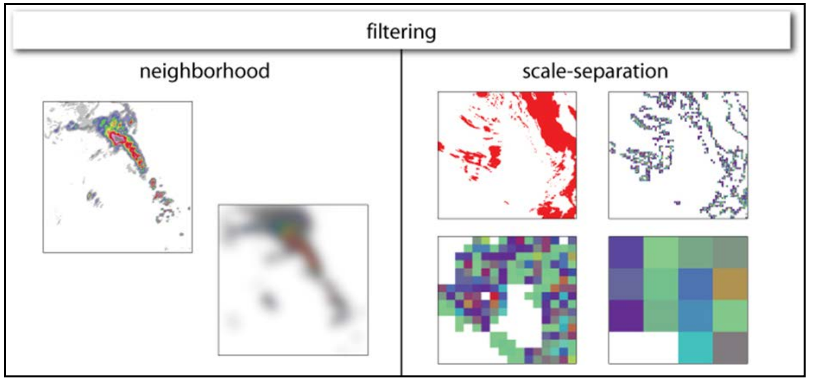
Method intercomparison projects

- First International Intercomparison Project (2006-2011)
  - Participants applied methods to a variety of cases:
    - Geometric
    - Real precipitation cases from US Midwest
    - Modified real cases (e.g., known displacements)
  - Summarized in several publications in *Weather and Forecasting* and *Bulletin of the AMS* (see reference list at http://www.rap.ucar.edu/projects/icp/index.html)
- Second International Intercomparison Project: MesoVICT ("Mesoscale Verification in Complex Terrain")
  - Focus:
    - Precipitation and wind in complex terrain (Alps region)
    - Ensemble forecasts and analyses
  - Kick-off workshop Oct 2-3, 2014 in Vienna, Austria
  - Web site: http://www.ral.ucar.edu/projects/icp/

Method strengths and limitations: Filtering methods

Strengths

- 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



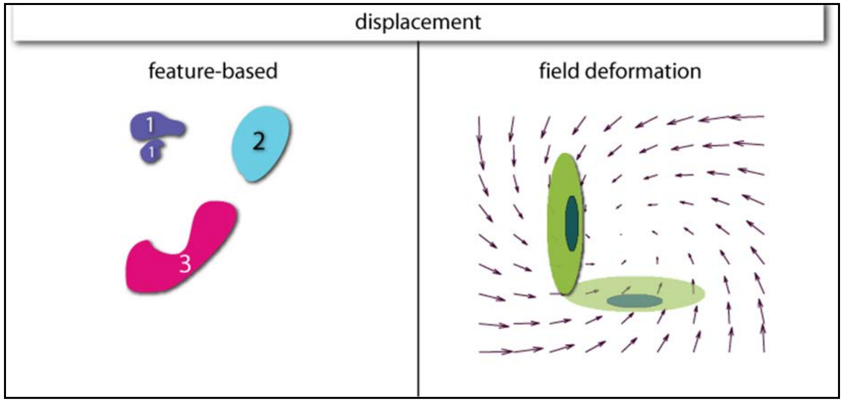
Limitations

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

Method strengths and limitations: Displacement methods

Strengths

- Features-based
  - Gives credit for close forecast
  - Measures displacement, structure
    - Provides diagnostic information
- Field-deformation
  - Gives credit for a close forecast
  - Can be combined with a field comparison significance test



Limitations

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

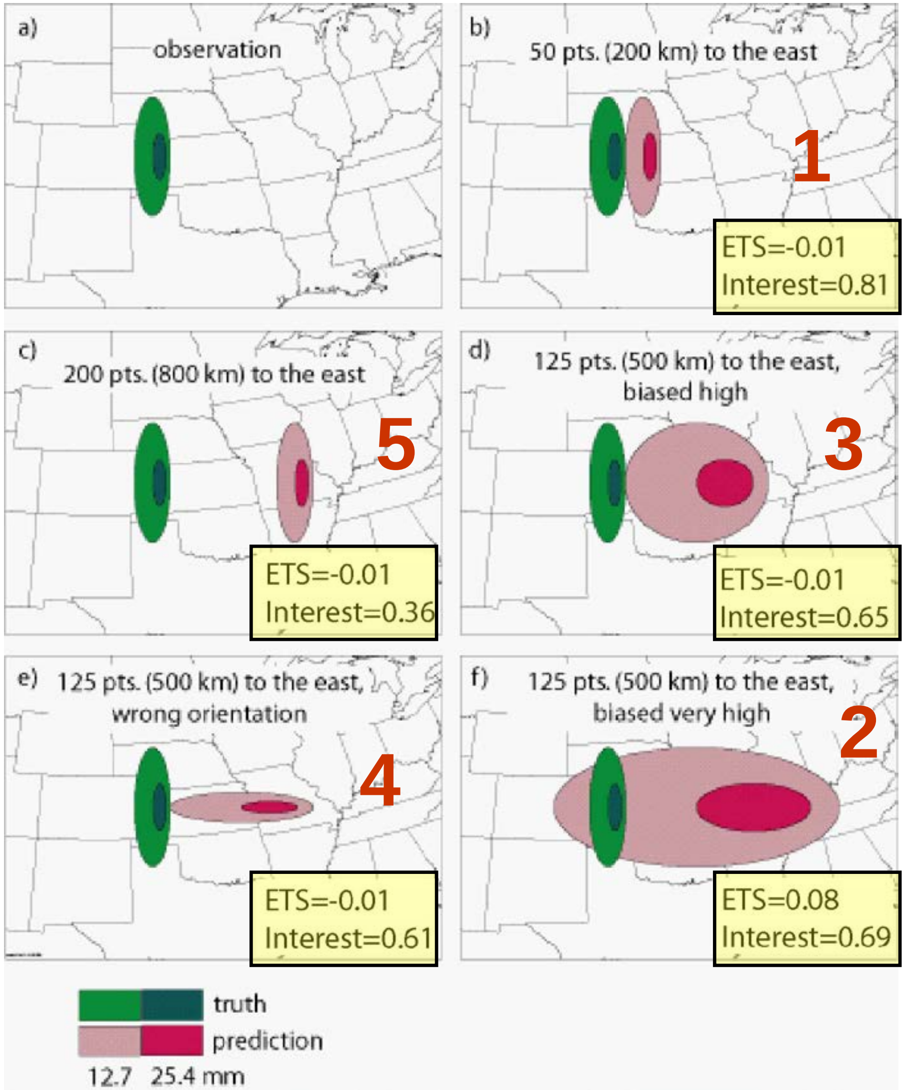
What do the new methods measure?

	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
	Scale Separation	Yes	Yes	No	Sensitive, but no direct information	Yes	Yes
Displacement	**Field deformation**	No	No	No	Yes	Yes	No
	Features-based	Indirectly	No	Yes	Yes	Yes	Yes, based on features rather than gridpoints

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

<u>Example:</u>

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

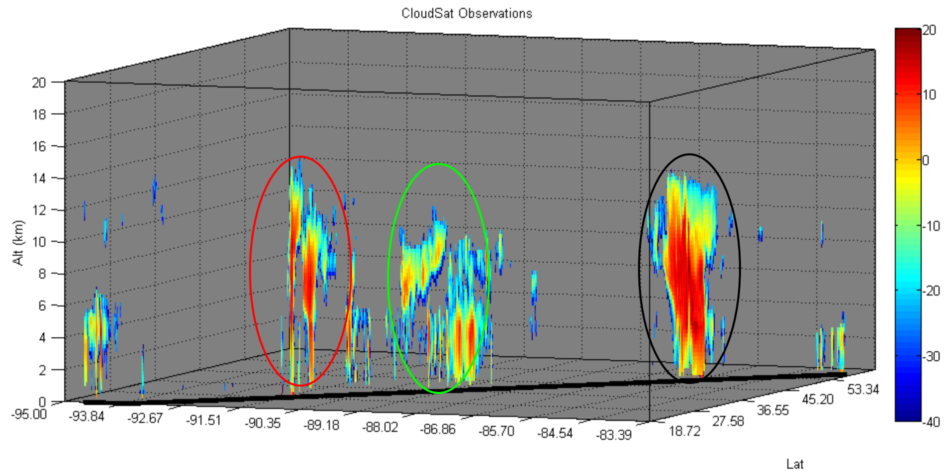


Application to other fields

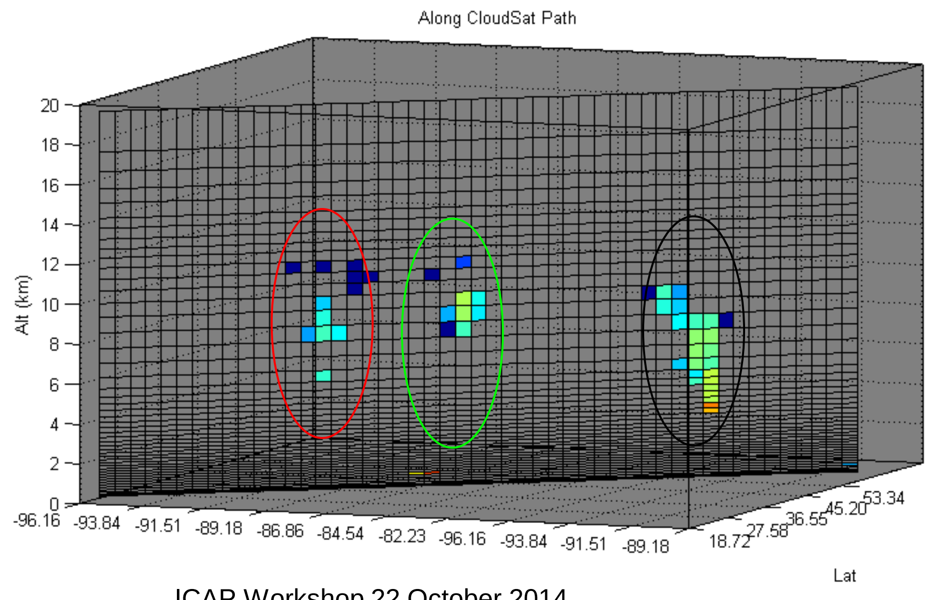
- Methods have been commonly applied to precipitation and reflectivity
- New applications
  - Wind
  - Cloud analysis
  - Vertical cloud profile
  - Satellite estimates of precipitation
  - Tropical cyclone structure
  - Ensemble
- Time dimension can also be included (see Fowler presentation)

Cloud-Sat Object-based Comparison: Along Track

CPR reflectivity

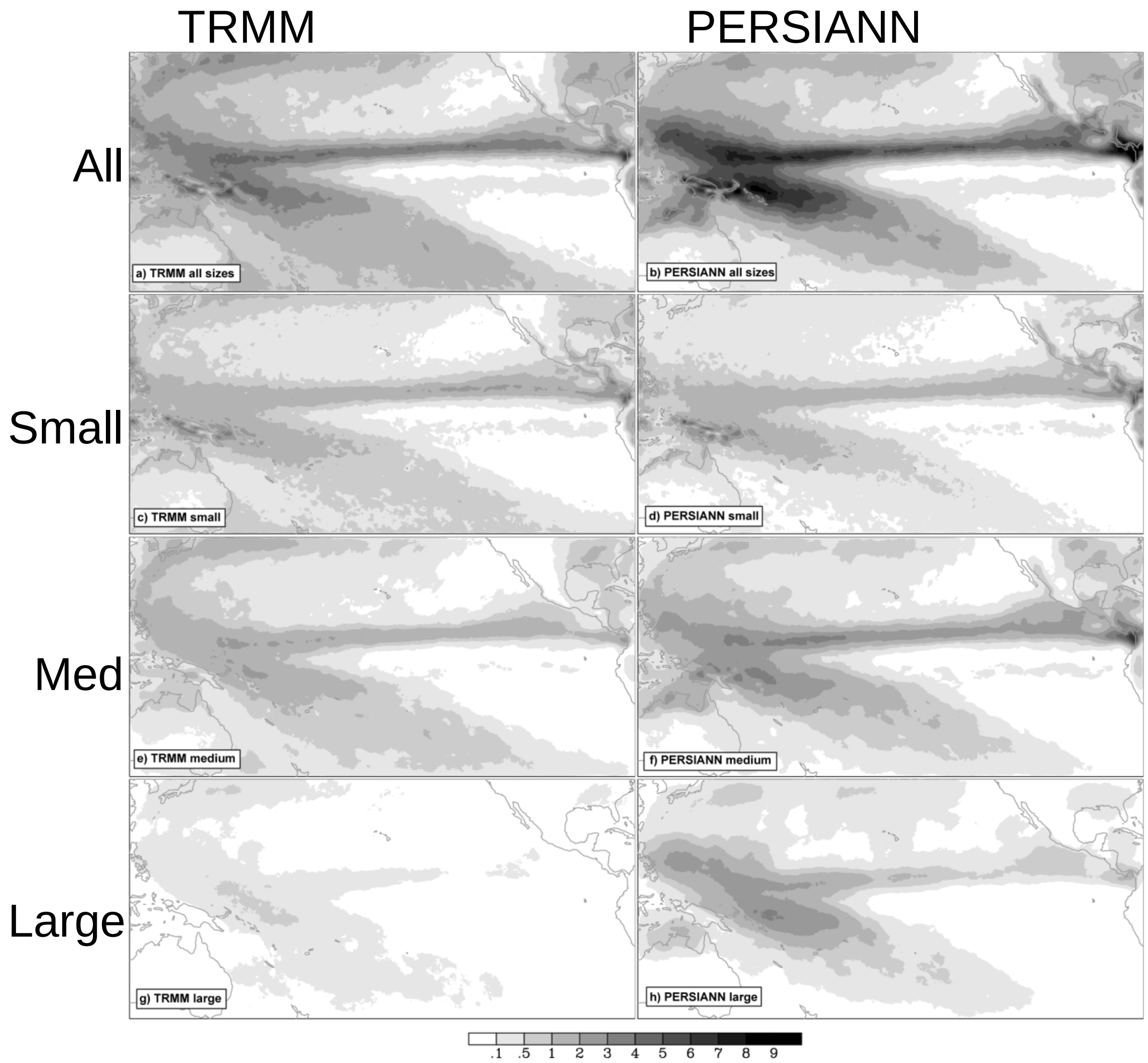


RUC reflectivity



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 a wide variety of 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 NCAR/RAL/JNTP
- 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