A closer look at methods for evaluation of extreme and spatial forecasts

Barbara Brown

Weather Systems and Assessment Program
Research Applications Laboratory, NCAR
Boulder, Colorado USA
bgb@ucar.edu

14 March 2018
CWB, Taipei, Taiwan

With acknowledgments to Eric Gilleland, Tara Jensen, Beth Ebert, Caspar Ammann, Tina Kalb, Randy Bullock

NCAR

Outline

- Methods for extremes
  - What are extremes and how are they different
  - Statistical analysis and representation of extremes
  - Categorical approaches
    - Extreme Dependency scores
- Spatial methods
  - Motivation
  - Method categories
  - Application to weather
  - Example applications to climate

EXTREMES

Why are special methods needed to evaluate extreme events?

Consider the MSE

- Forecast 1:
  - 0.9 cm of precipitation is forecast; 1 cm of precipitation occurs
  - Error is 0.1 cm (10%)
    - Contribution to MSE is 0.01
- Forecast 2:
  - 9 cm of precipitation is forecast;10 cm of precipitation occurs
  - Error is 1 cm (10%)
    - Contribution to MSE is 1
- Is this what we want to happen? The extreme precipitation is much more harshly penalized than the “easy” forecast – even for a pretty good forecast!

Extreme events

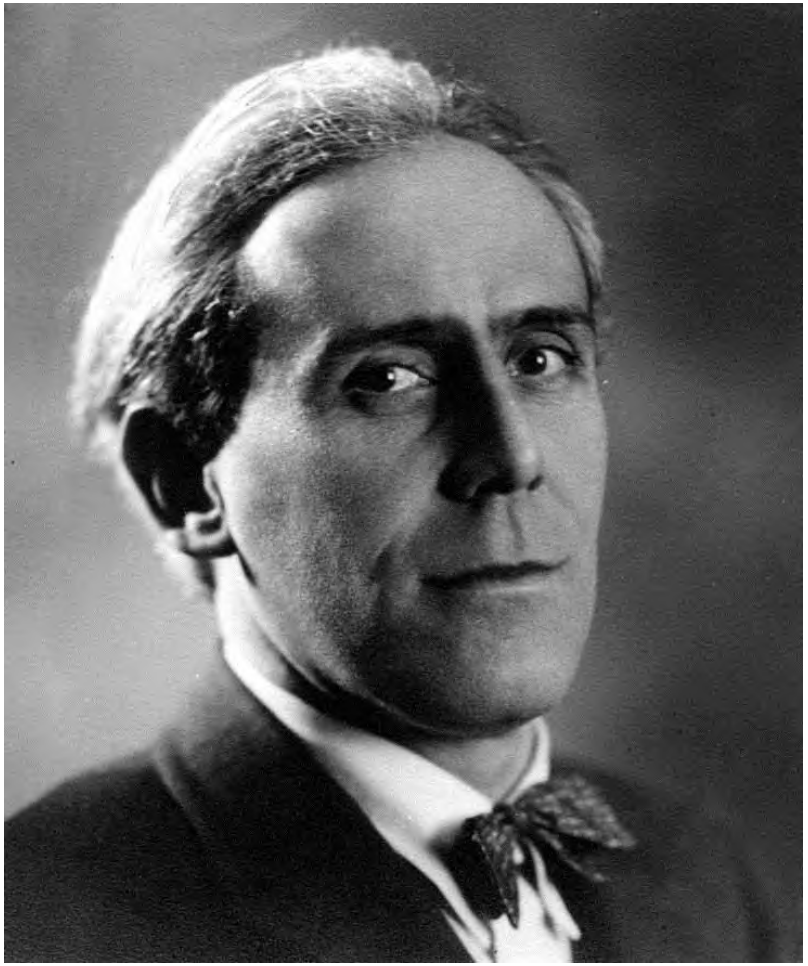
- "*Extreme*" weather often implies "*Rare*" or infrequent event – i.e., small samples
- Infrequent events (low "base rate" or climatology) often require special statistical treatment...
- May be difficult to observe
  - Greater observation error/uncertainty

Gare Montparnasse, 1895

Extreme Value Theory/Analysis

"Il est impossible que l'improbable n'arrive jamais"

--Emil Gumbel



"It is impossible that the improbable will never happen"

--Emil Gumbel

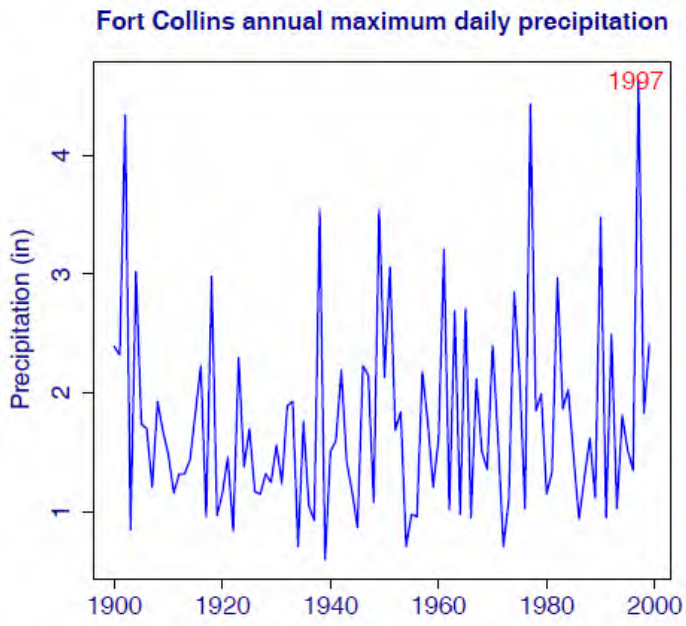
Other pioneers:

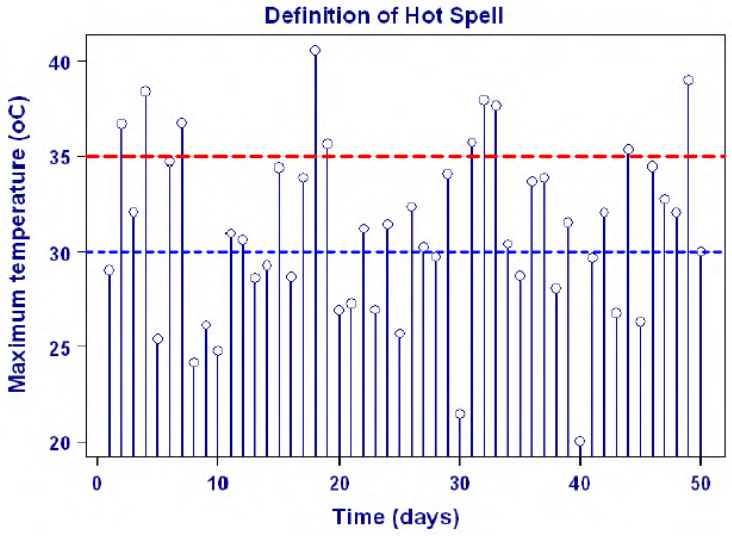
- Fisher
- Tippett
- Weibull
- Pareto

Extreme value theory

Focuses on

- The distributions of *maximums* (or minimums)
  - *Example*: The maximum yearly 24-h precipitation amount
- The *frequency* of extreme events
  - *Example*: The number of times daily maximum temperature exceeds 35C



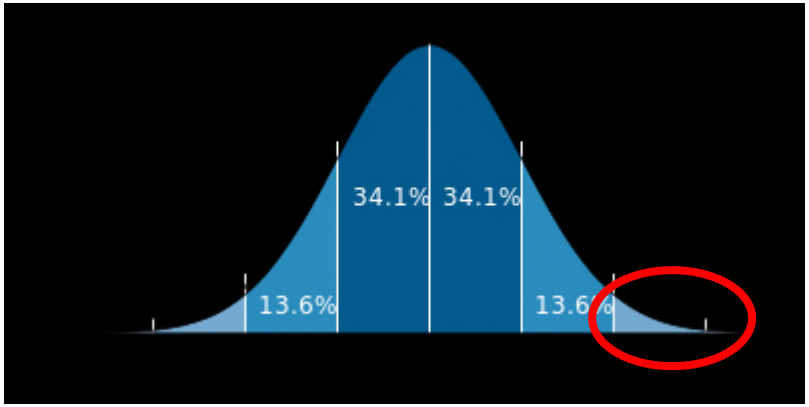


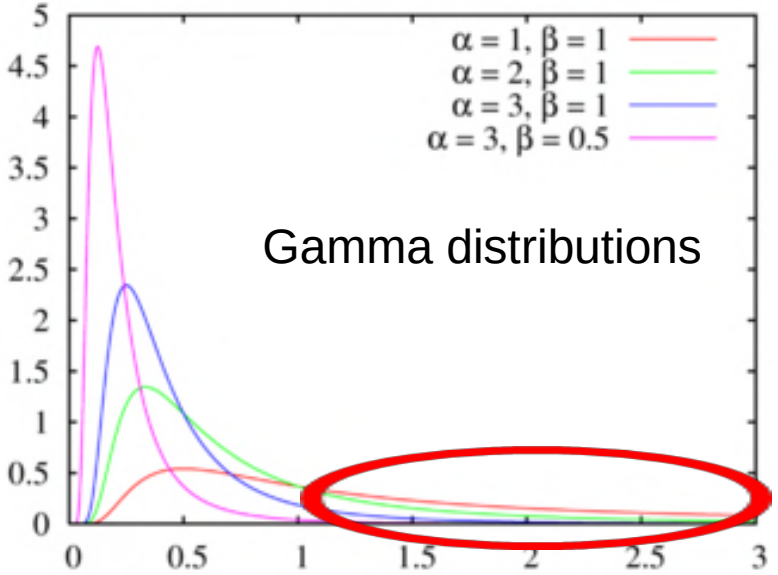
Credit: RW Katz

Extreme value distributions

- Extreme value theory concerns the *tails* of the underlying distribution
- The distribution of these extreme values is *not the same* as the "parent" distribution
- The "*Generalized Extreme Value (GEV)*" distribution has been developed to describe these distributions

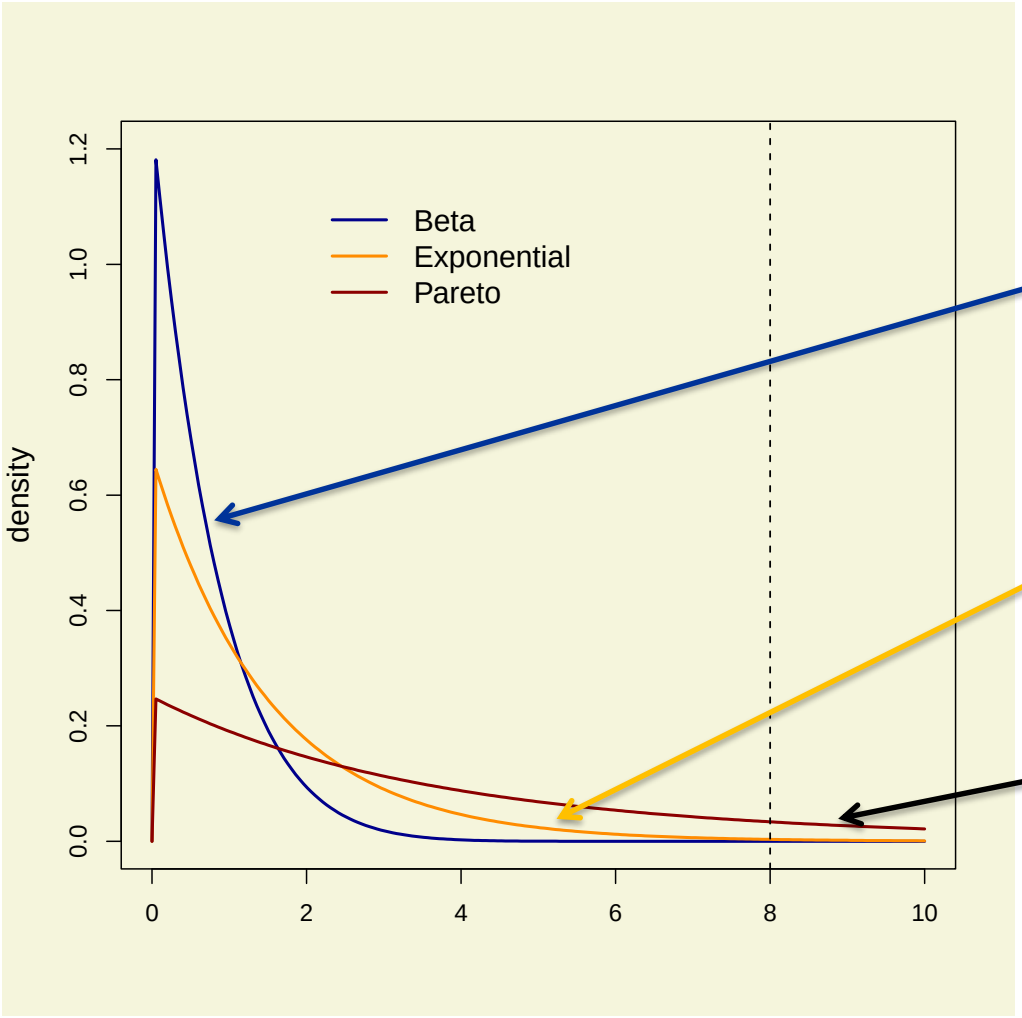
Normal distribution





Extreme Value Analysis: Peaks over thresholds

Three types of Extreme Value df's



Beta df; bounded upper tail

Exponential; light tail

Pareto; heavy tail

Choices for verification of extremes

- Need methods that provide information about the extremes
  - Measures like MSE, MAE emphasize the “middle” of the distribution
  - Many measures highly penalize errors in extremes
- Approaches to consider:
  - Categorical scores
  - Extreme dependency scores
  - Probabilistic and spatial approaches
  - “Alternative” scores (e.g., spatial methods)

Categorical scores

Advantages

- Allow “user” to select thresholds for events of interest
  - Threshold can be a (very) large (or small) value to represent relevant events
- Not sensitive to sizes of errors

Disadvantages

- Dependent on sample size (to capture enough events)
- Sensitive to over-forecasting (large biases)
- Not sensitive to sizes of errors
- “Degenerate” for very extreme events (more about this later)

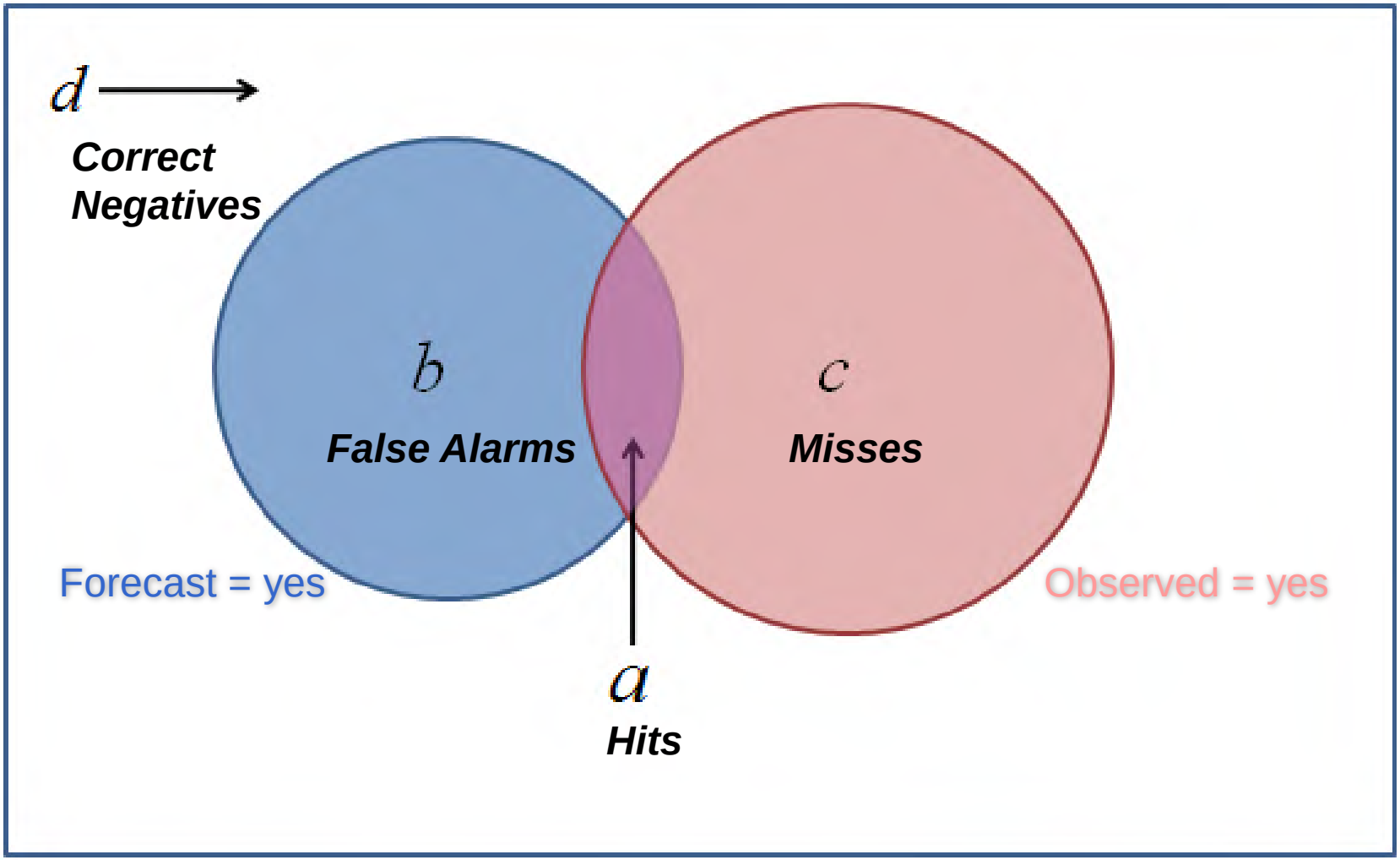
2 x 2 Contingency Table

		Observed		
Forecast		Yes	No	Total
	Yes	Hit	False Alarm	Forecast Yes
	No	Miss	Correct Negative	Forecast No
	Total	Obs. Yes	Obs. No	Total

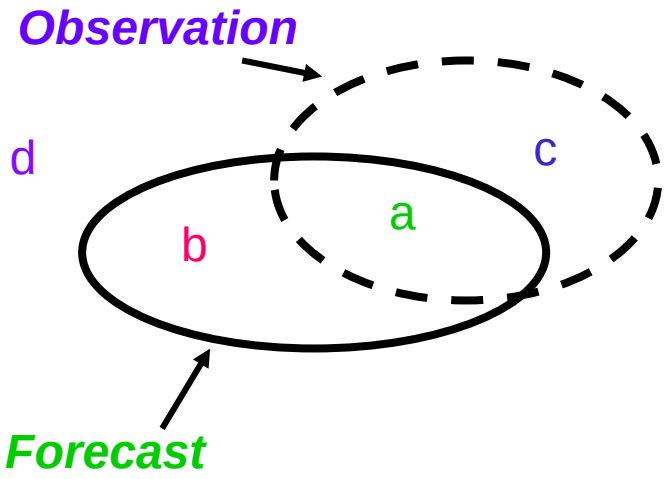
Example:

$$\% \text{ Correct} = 100 \times \frac{(\text{Hits} + \text{Correct Negs})}{\text{Total}}$$

Alternative Perspective on Contingency Table



"Standard" Verification Measures (Yes/No forecasts)



- a = Hits
- c = Misses
- b = False Alarms
- POD = a / (a + c)
- POFD = b / (b + d)
- FAR = b / (a + b)
- Bias = (a + b) / (a + c)

- **Accuracy = (a+d) / (a+b+c+d)**
  Measures overall % correct
- **CSI = a / (a + b + c)**
  Measures "relative accuracy"
- **H-K = POD + POFD -1**
  Measures "discrimination" between Yes and No observations
- **POD (PODy)**
  Measures proportion of observed area that is correctly forecast to be "Yes"
- **POFD (PODn)**
  ***Measures proportion of area that is correctly forecast to* be "No"**
- **FAR**
  Measures proportion of forecast convective area that is incorrect
- **Bias**
  Measures the extent of over- or under- forecasting
- **Skill scores (Heidke, Gilbert/ETS)**
  Measure the improvement in Accuracy and CSI, respectively over what's expected by chance

Categorical methods originally designed for extremes

- Heidke skill Score (HSS)
  Accuracy corrected for number correct expected by chance

$$\mathrm{HSS} = \frac{a + d - C_H}{a + b + c + d - C_H}$$

where $C_H = \frac{(a+b)(a+c) + (b+d)(c+d)}{n}$

- Critical Success Index (CSI)
  Accuracy, ignoring Correct Negatives (d)

$$\mathrm{CSI} = \frac{a}{a + b + c}$$

- Gilbert Skill Score (GSS)
  CSI corrected for number of hits expected by chance

$$\mathrm{GSS} = \frac{a - C_G}{a + b + c - C_G}$$

where $C_G = \frac{(a+b)(a+c)}{n}$

Finley revisited…

	Obs. Yes	Obs. No	Sum
Fcst.Yes	28	72	100
Fcst. No	23	2680	2703
Sum	51	2752	2803

$$\text{POD} = \frac{28}{51} = 0.55$$

$$\text{FAR} = \frac{72}{100} = 0.72$$

$$\text{Bias} = \frac{100}{51} = 1.96$$

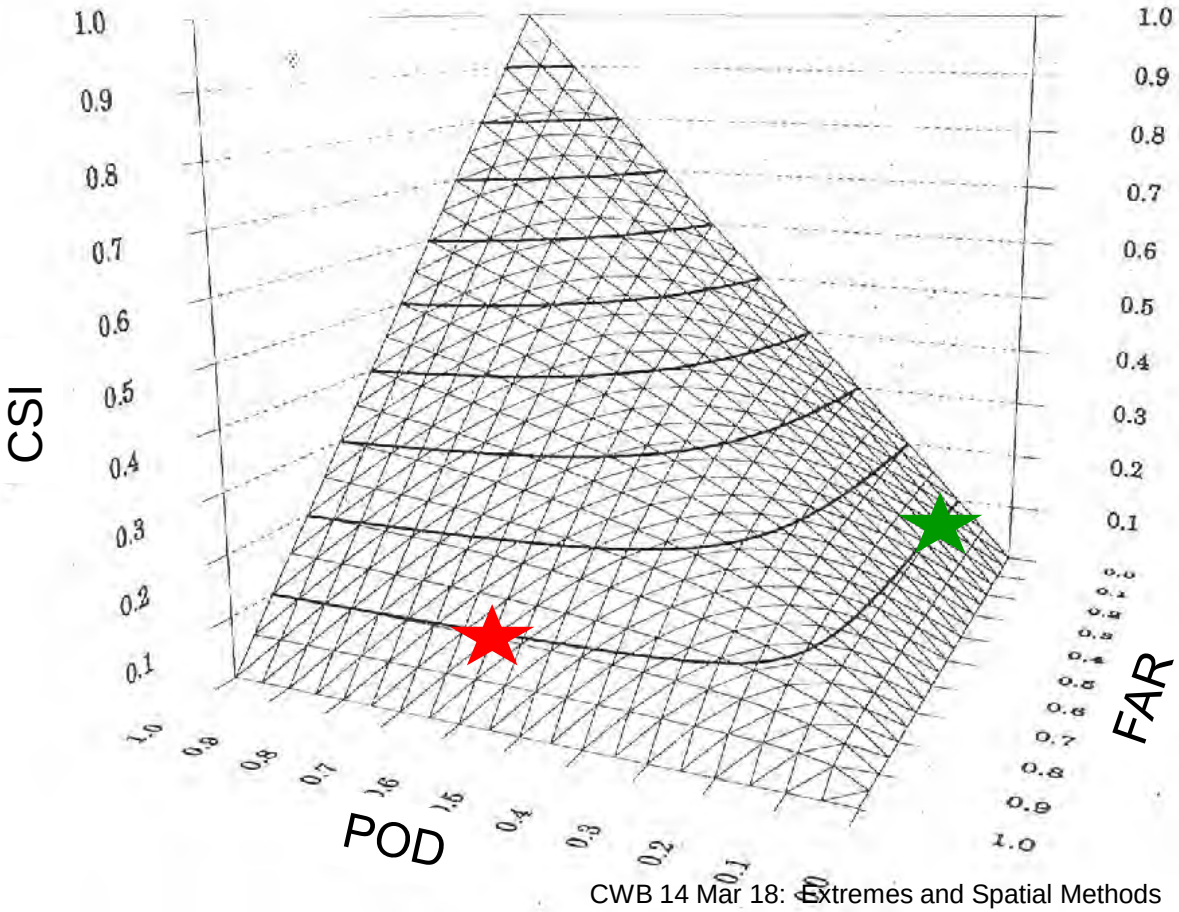
$$\text{CSI} = \frac{28}{28+72+23} = 0.23$$

$$\text{ETS} = \frac{28-1.8}{123-1.8} = 0.11$$

$$\text{HSS} = \frac{28+2680-2656}{2803-2656} = 0.35$$

Relationships among contingency table scores

- CSI was designed to focus on extreme/rare events
- CSI is a *nonlinear* function of POD and FAR
- CSI depends on base rate (event frequency) and Bias



$$\mathrm{CSI} = \frac{a}{a+b+c}$$

$$\mathrm{CSI} = \frac{1}{\frac{1}{\mathrm{POD}} + \frac{1}{1-\mathrm{FAR}} - 1}$$

Very different combinations of FAR and POD lead to the same CSI value

Precipitation Performance Diagram

All on same plot

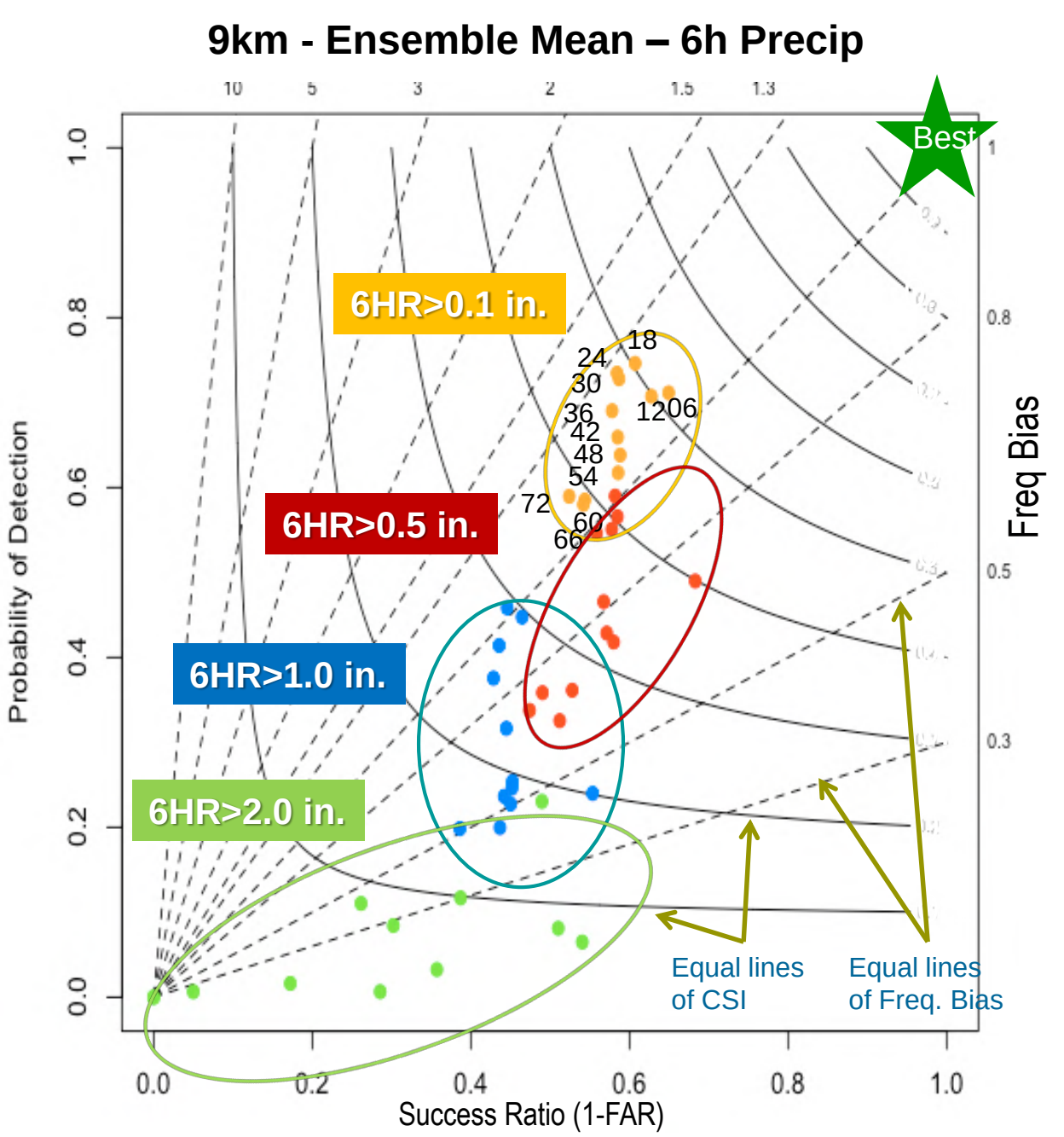
- POD
- 1-FAR (aka Success Ratio)
- CSI
- Freq Bias

Dots: Scores Aggregated Over Lead Time

Colors: Different Thresholds

Here we see:

- Decreasing skill with higher thresholds even with multiple metrics
- Highest skill at 18-24h leads



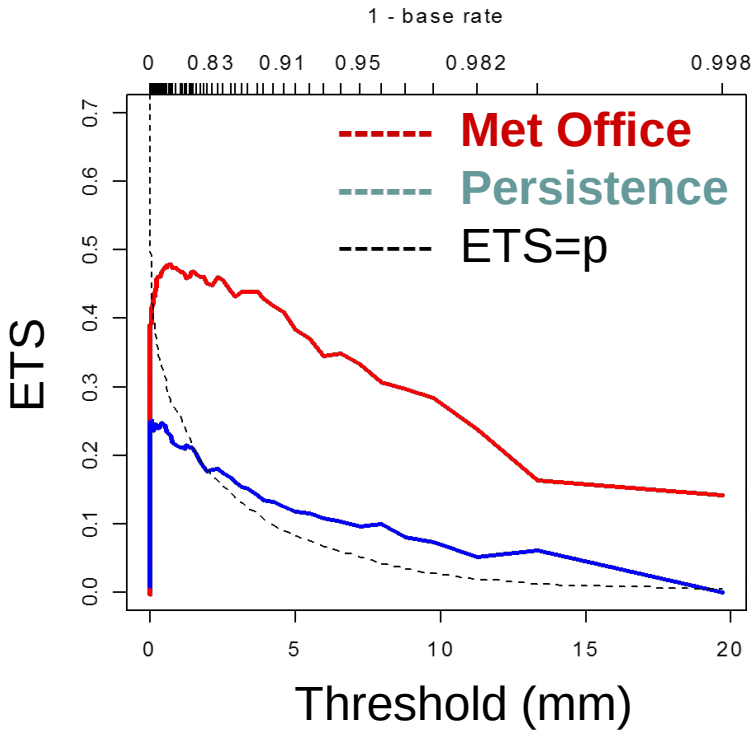
Problem with using traditional contingency table approach

Stephenson and Ferro:

All of the standard measures are "degenerate" for large values…

That is, they tend to a meaningless number (e.g., 0) as the base rate (climatology) gets small – as the event becomes more extreme

Result: It looks like forecasting extremes is impossible



ETS example (Stephenson)

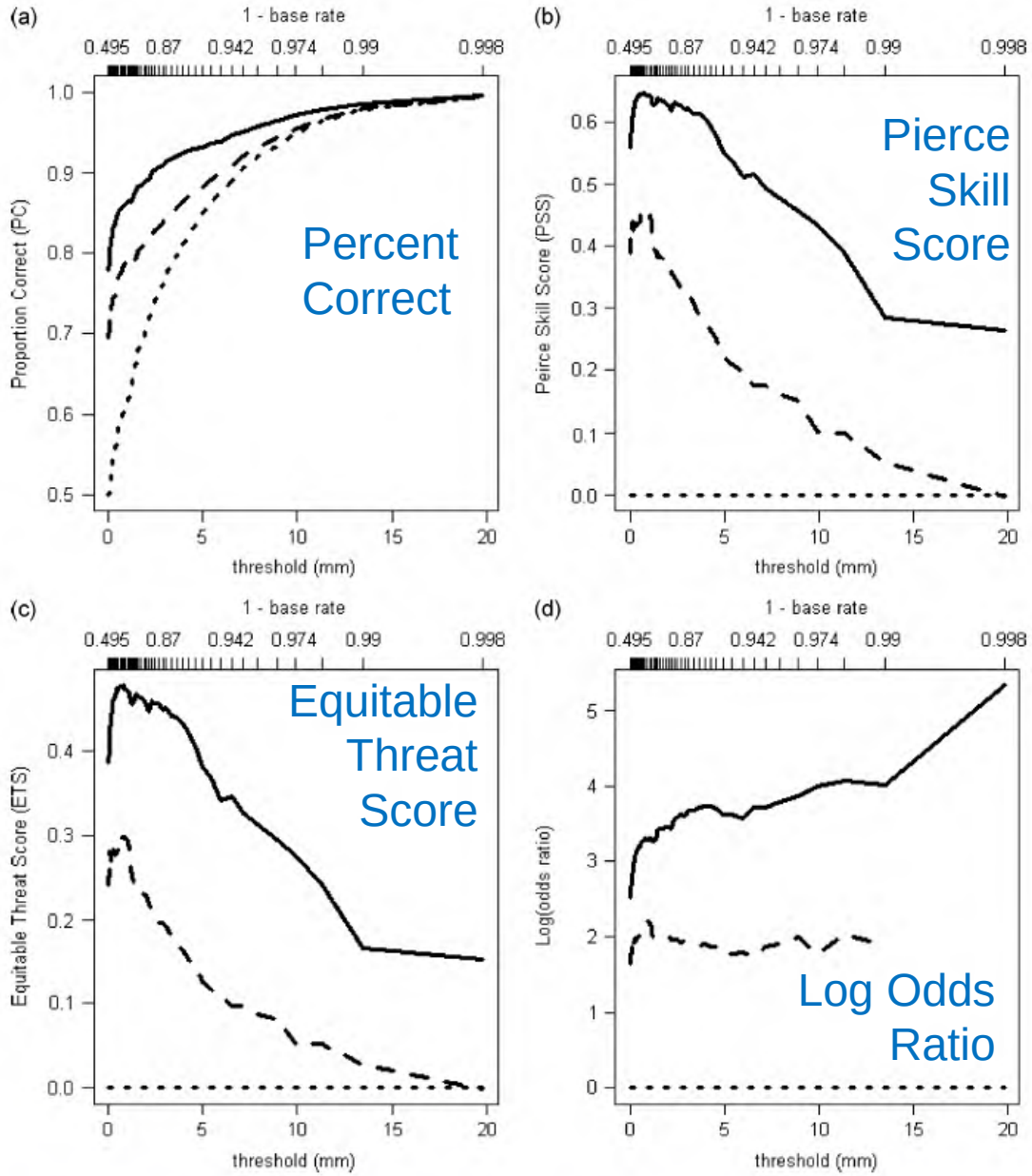
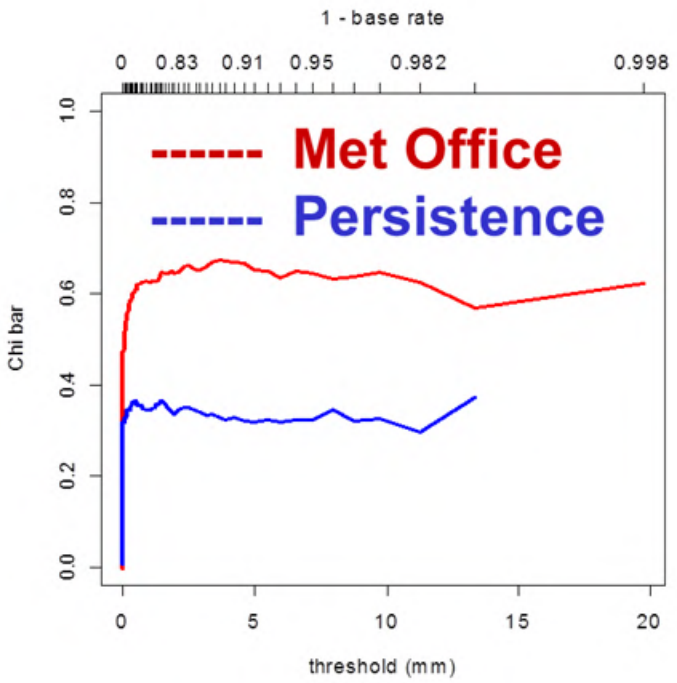


Figure 4. Scores *versus* threshold: (a) proportion correct PC, (b) the Peirce skill score (PSS), (c) the equitable threat score (ETS), and (d) logarithm of the OR. Solid line denotes Met Office forecasts, dashed line denotes 6-h-lead persistence forecasts, and the dotted line denotes random forecasts.

Stephenson et al. 2008

New measures for extremes

- "***Extreme dependency scores***" developed starting in 2008 by Stephenson, Ferro, Hogan, and others
- Based on asymptotic behavior of score with decreasing base rate
- All based on contingency table counts
- Catalog of measures
  - EDS – Extreme Dependency score
    - Found to be subject to hedging (over-forecasting)
  - SEDS – Symmetric EDS
    - Dependent on base rate
  - SEDI – Symmetric Extremal Dependency Index
    - Closer to a good score
    - Being used in practice in some places



(See Chapter 3 in Jolliffe and Stephenson 2012; Ferro and Stephenson 2011; *Weather and Forecasting*)

Extreme dependency score example

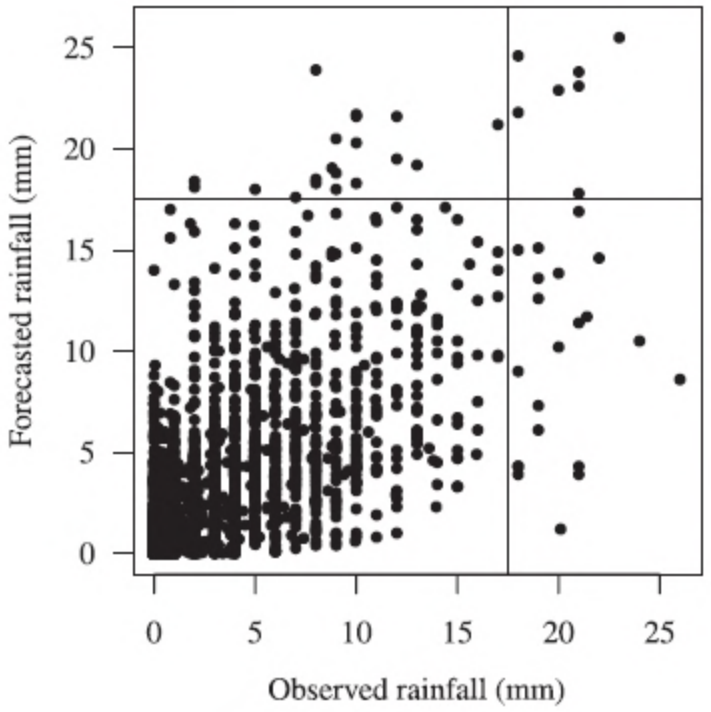
Symmetric Extreme Dependency Index (SEDI):

$$SEDI = \frac{\ln F - \ln H + \ln(1-H) - \ln(1-F)}{\ln F + \ln H + \ln(1-H) + \ln(1-F)}$$

Where

$$H = \text{Hit Rate} = \frac{a}{a+c} \qquad F = \text{False Alarm Rate} = \frac{b}{b+d}$$

EDS Example



6-h rainfall (Eskdalemuir)

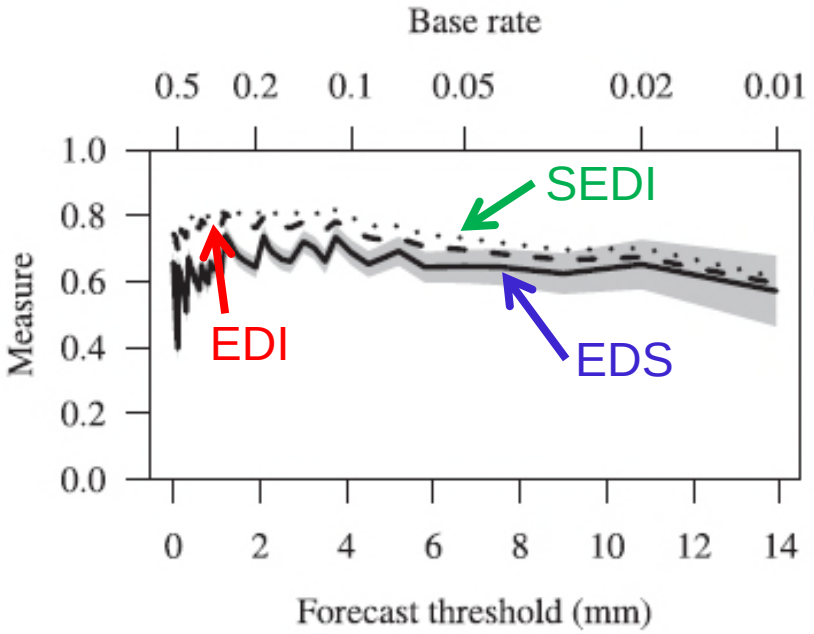


TABLE 6. Properties of five verification measures.

	ETS	EDS	SEDS	EDI	SEDI
Nondegenerate limit	×	✓	✓	✓	✓
Base-rate independent	×	×	×	✓	✓
Nontrivial to hedge	✓	×	✓	✓	✓
Regular	×	×	×	✓	✓
Fixed range [−1, 1]	×	×	×	✓	✓
Asymptotically equitable	✓	×	✓	✓	✓
Meaningful origin	✓	×	✓	✓	✓
Complement symmetric	✓	×	×	×	✓
Transpose symmetric	✓	×	✓	×	×

From Ferro and Stephenson 2011 (*Wx and Forecasting*)

Measure	Attribute evaluated	Comments
Probability forecasts		
Brier score	Accuracy	Based on squared error
Resolution	Resolution (resolving different categories)	Compares forecast category climatologies to overall climatology
Reliability	Calibration	
Skill score	Skill	Skill involves *comparison* of forecasts
Sharpness measure	Sharpness	Only considers distribution of forecasts
ROC	Discrimination	Ignores calibration
C/L Value	Value	Ignores calibration
Ensemble distribution		
Rank histogram	Calibration	Can be misleading
Spread-skill	Calibration	Difficult to achieve
CRPS	Accuracy	Squared difference between forecast and observed distributions Analogous to MAE in limit
log p score	Accuracy (IGN = $-\log_2 p$)	Local score, rewards for correct category; infinite if observed category has 0 density

Benefits and issues with probabilistic forecasts

Benefits

- Direct focus on event of interest
  - Probabilities associated with specific events
  - Clear identification of what is important
  - Recognition of uncertainty in forecasts
- Clear approaches for evaluation

Issues

- Limited range of probabilities
- Need large samples
- Observation uncertainties make reliability information questionable

SPATIAL METHODS

Spatial fields

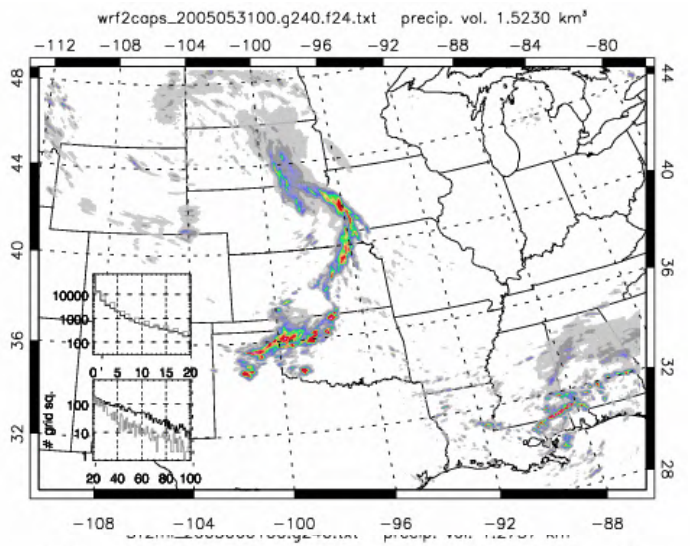
Weather and climate variables defined over spatial domains have **coherent spatial structure and features**

Traditional methods ignore this structure

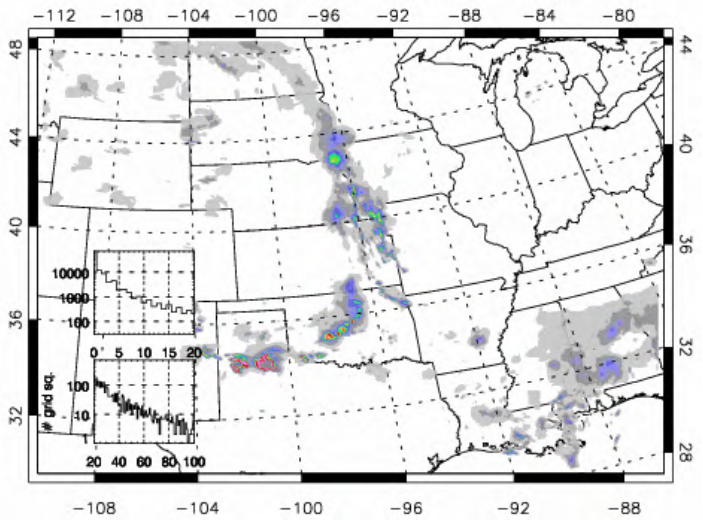
<u>Goal</u>: Define and compare spatial areas of interest

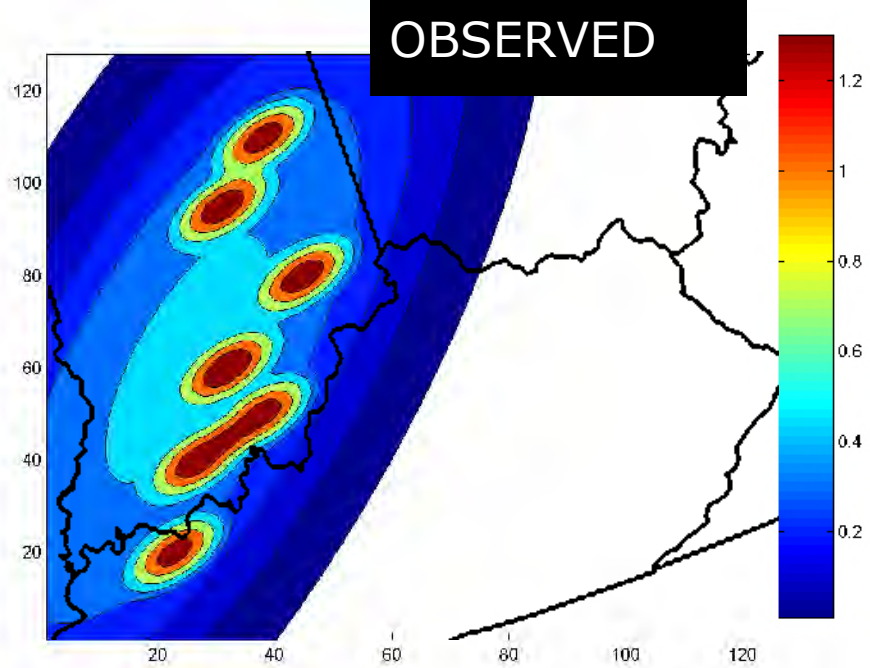
Alternative to treating forecasts as a collection of points

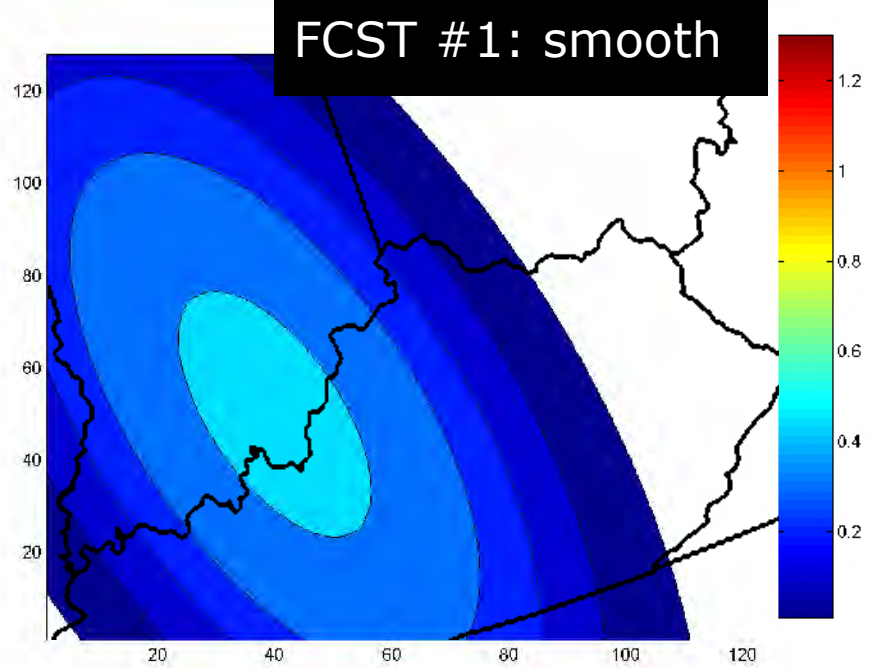
WRF model

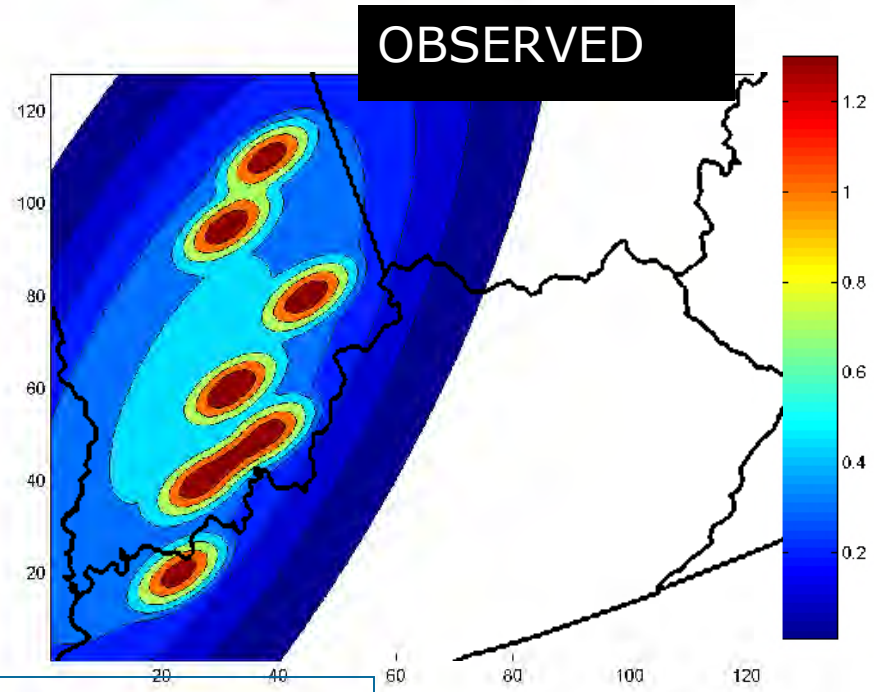


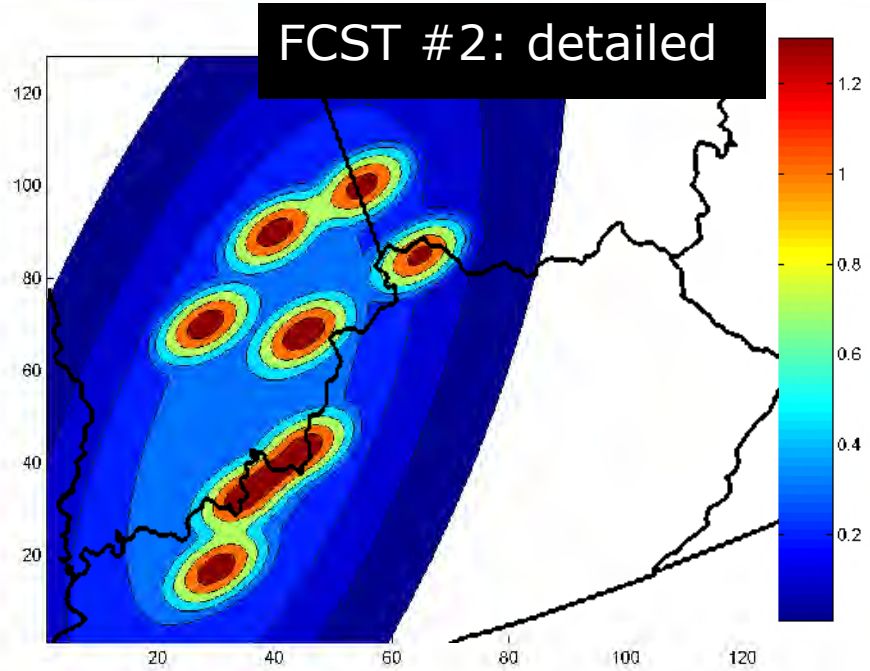
Stage II radar











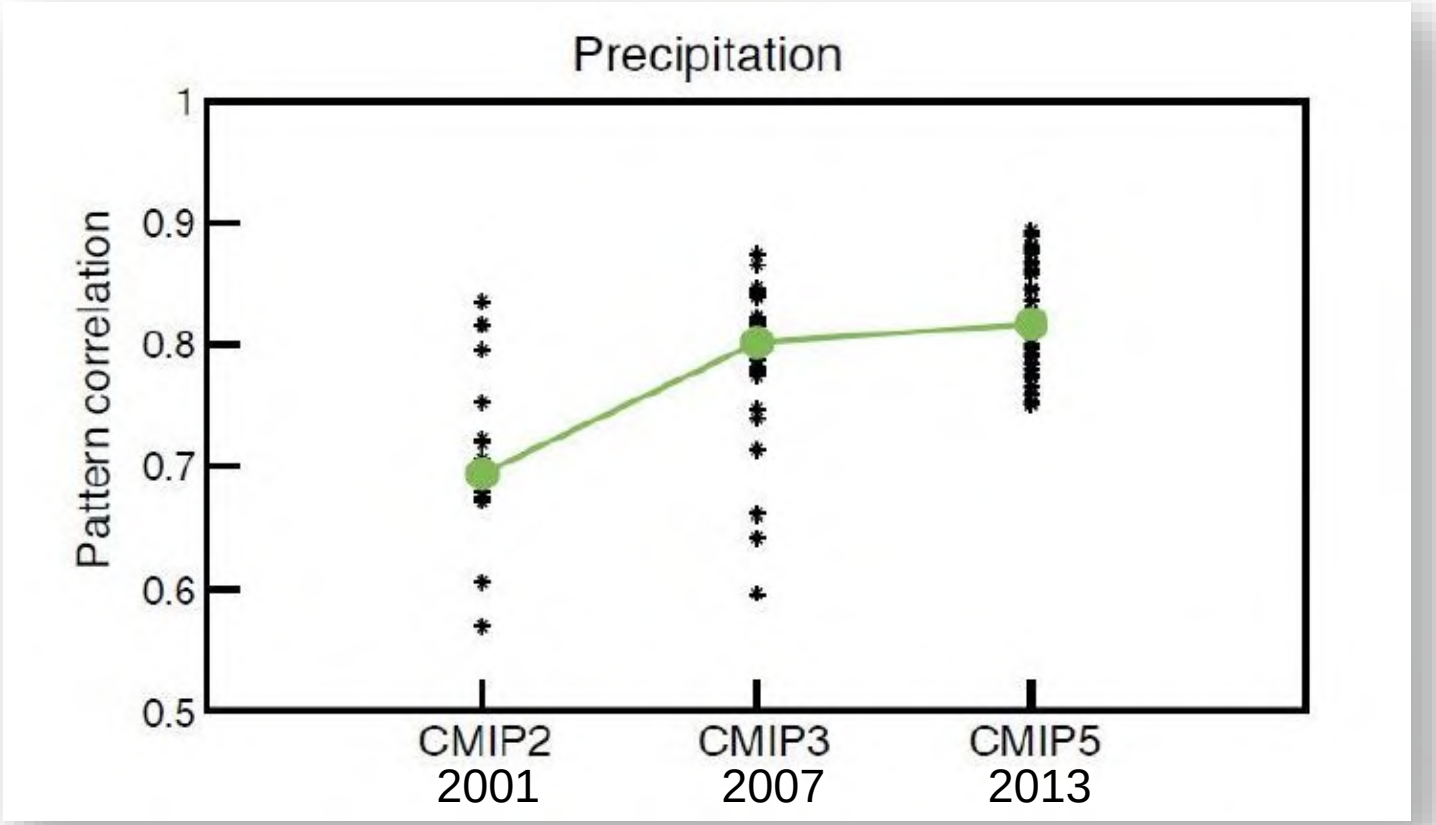
From Baldwin 2002

"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

Validation: Skill of Models



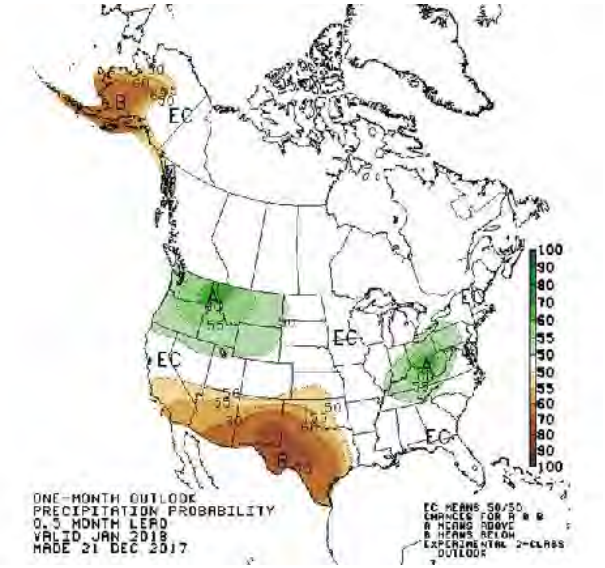
IPCC Model “Spatial Skill”: Pattern Correlations

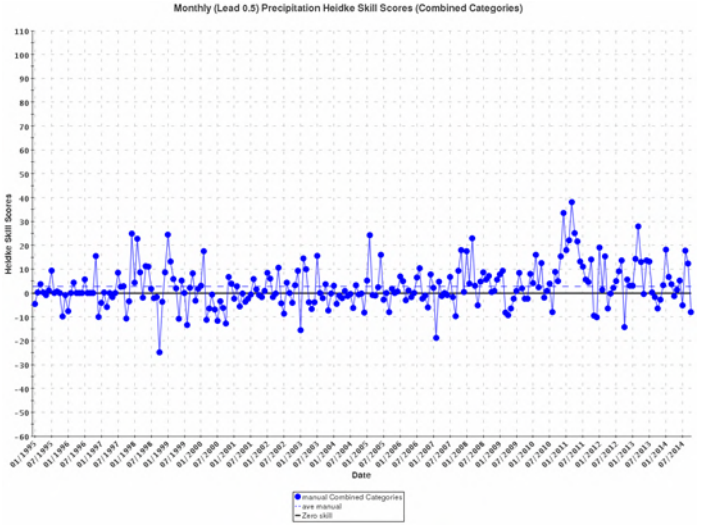
Spatial verification approach(es)

Some key questions for evaluation of S2S and climate models:

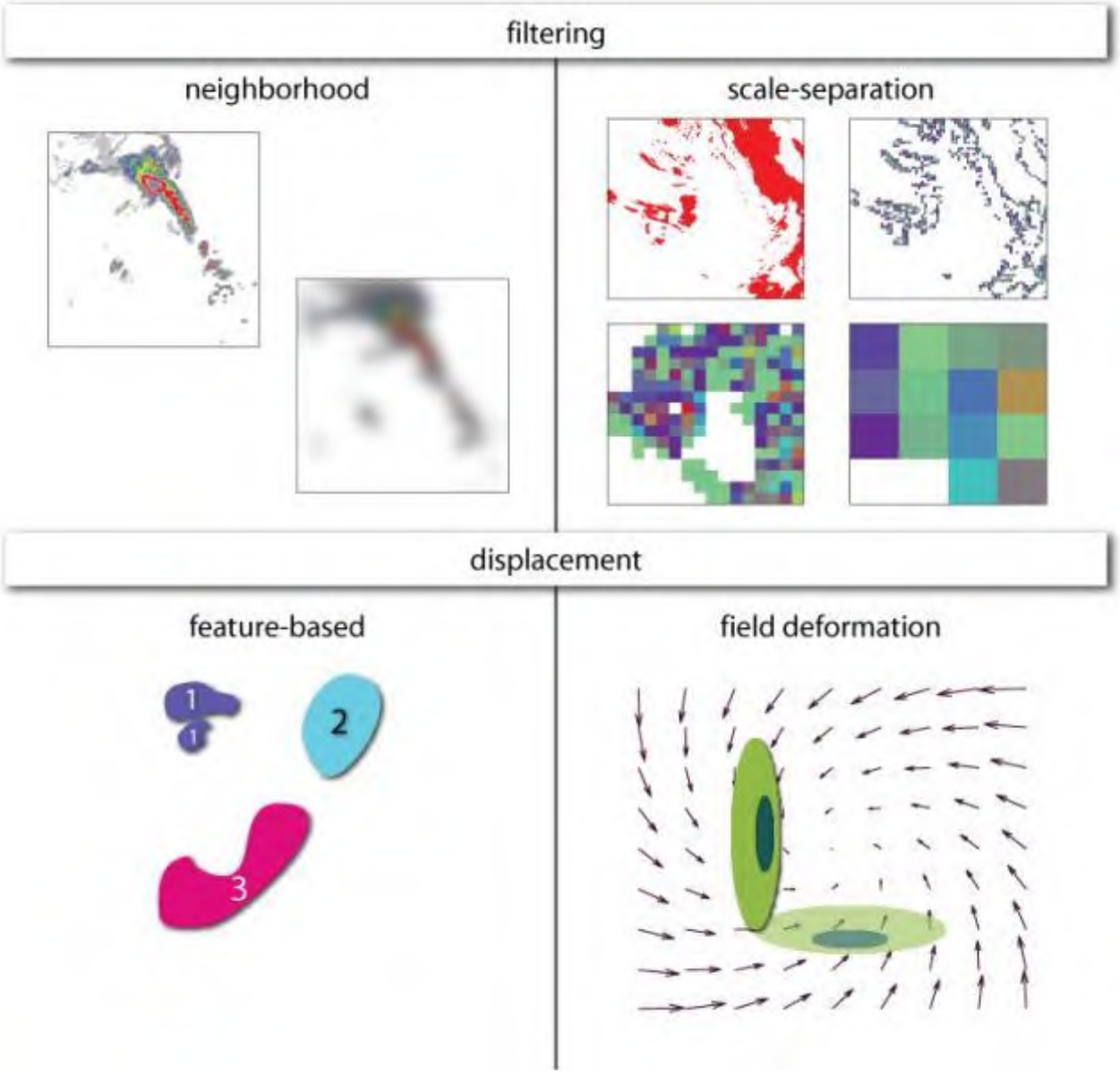
- How well does a model
- … reproduce S2S/climate characteristics?
- … represent spatial and temporal variations?
- … identify good and bad aspects of predictions?

Goal: Expand climate/S2S model evaluation "toolkit" to include spatial methods currently being applied for weather predictions





Spatial Method Categories



New spatial verification approaches

Neighborhood

Successive smoothing of forecasts/obs

Gives credit to "close" forecasts

Scale separation

Measure scale-dependent error

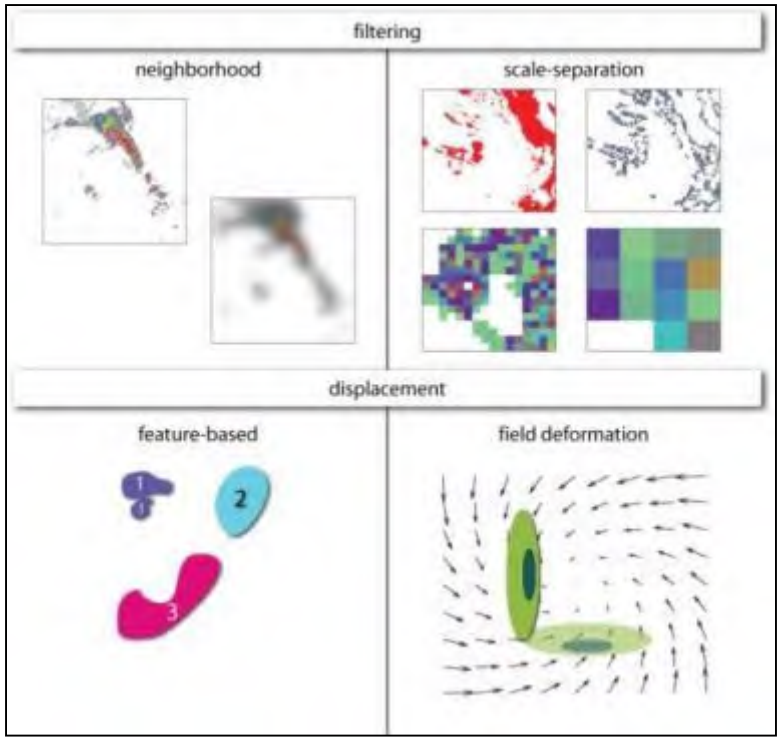
Field deformation

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

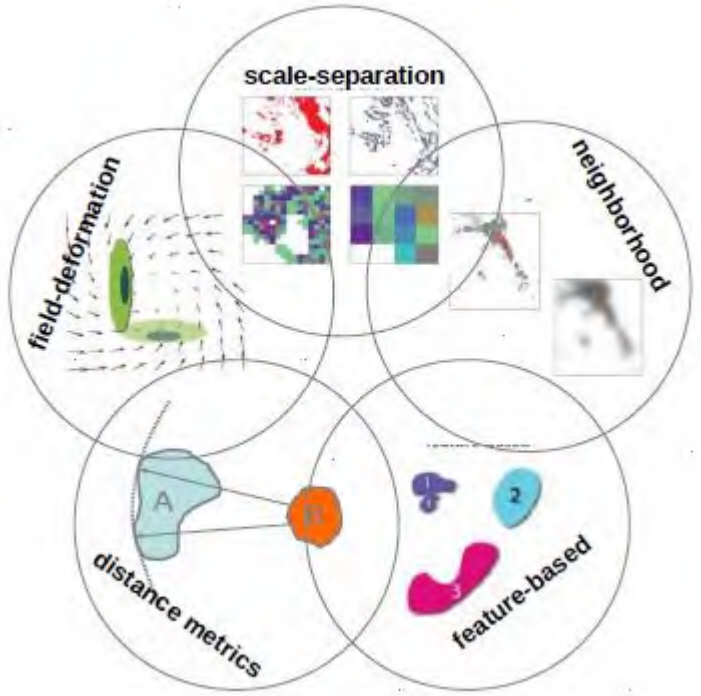
Evaluate attributes of identifiable features



5th category: Distance metrics

Distance metrics:

Measure the overall distance between a forecast field and an observation field



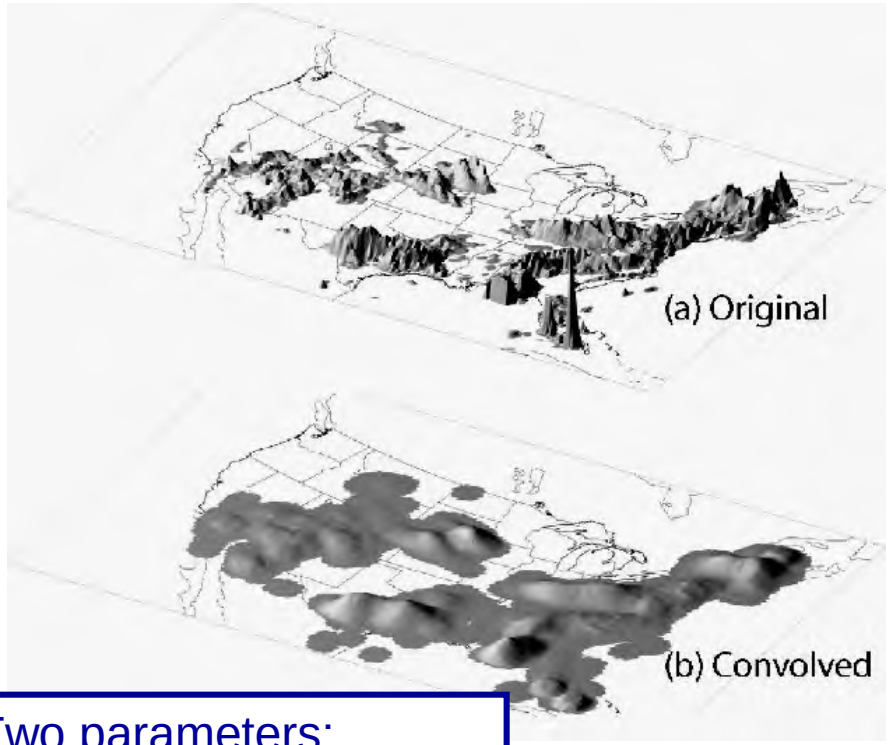
Dorninger et al. 2018; *BAMS*

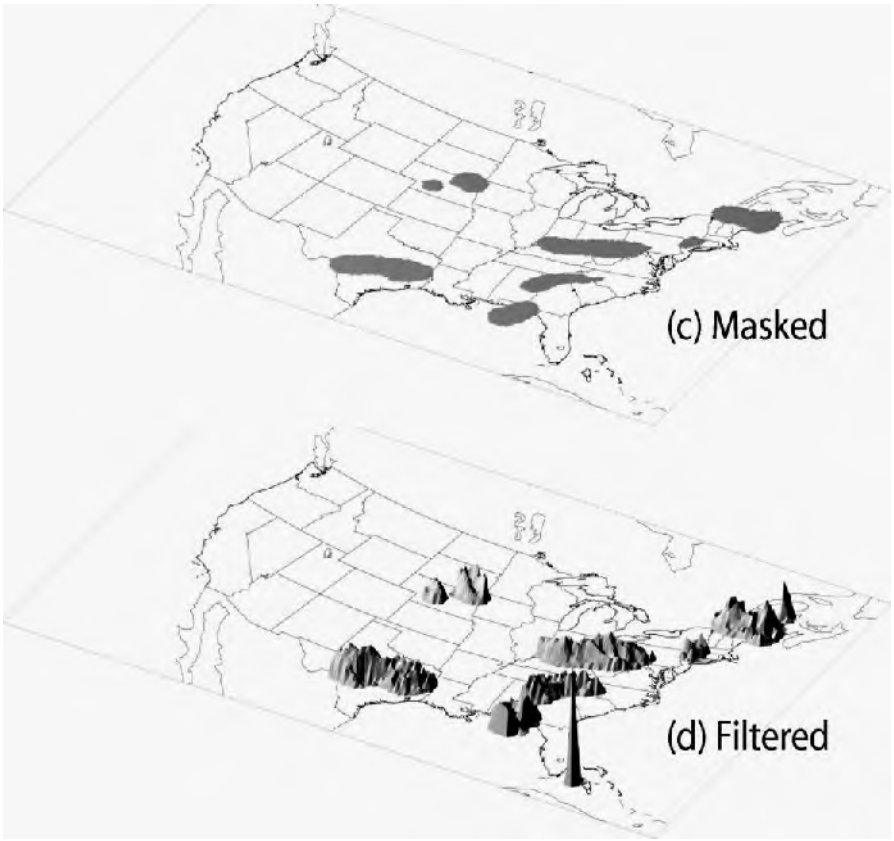
Commonly used spatial methods

- Neighborhood
  - Fractions skill score
    - Successive smoothing and comparison of forecast vs. observed coverage
    - Compare performance to scale (smoothing level)
- Distance metrics
  - Baddeley's Delta
  - Mean Error Distance (MED)
- Object-based
  - MODE: Method for Object-based Diagnostic Evaluation
  - CRA: Contiguous Rain Area (Ebert-McBride)
  - SAL: Structure-Amplitude-Location
- Field deformation
  - Image warping

MODE – Method for Object-based Diagnostic Evaluation

Davis et al., *MWR*, 2006





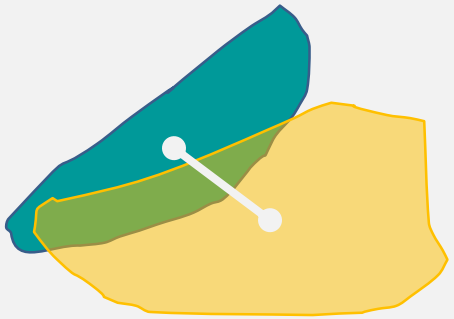
Two parameters:

1. Convolution radius
2. Threshold

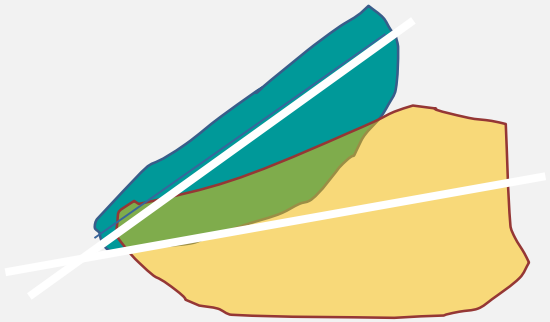
Attributes of Objects defined by MODE

Observed Field

Forecast Field

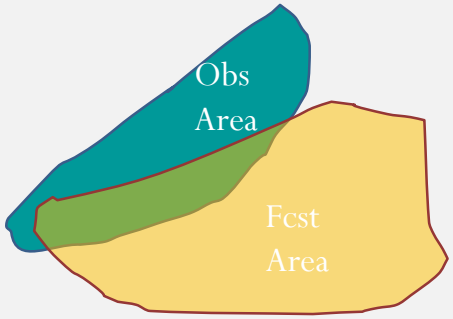


Centroid Distance: Provides a quantitative sense of spatial displacement.
Small is good



Axis Angle: Provides an objective measure of how well the objects are aligned. ***Small is good***

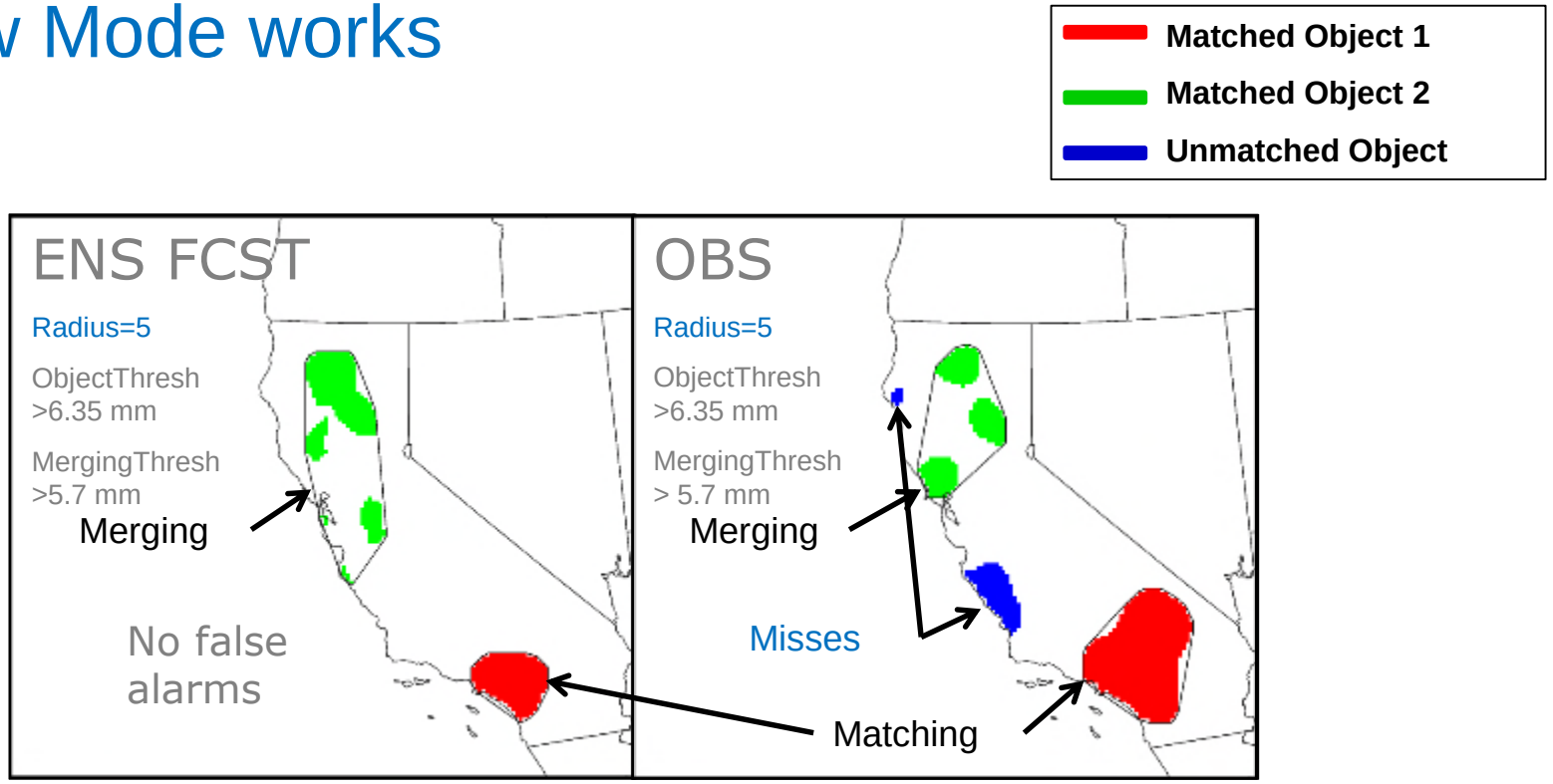
Area Ratio =
Fcst Area
Obs Area



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

Method for Object-based Diagnostic Evaluation (MODE)

How Mode works



Comparing objects can tell you things about your forecast like . . .

This:	Instead of this:
30% Too Big (area ratio=1.3)	POD = 0.35
Shifted west 1 km (centroid distance = 1km)	FAR = 0.72
Rotated 15° (angle diff = 15%)	CSI = 0.16
Peak Rain 1/2" too much (diff in 90th percentile of intensities = 0.5)	

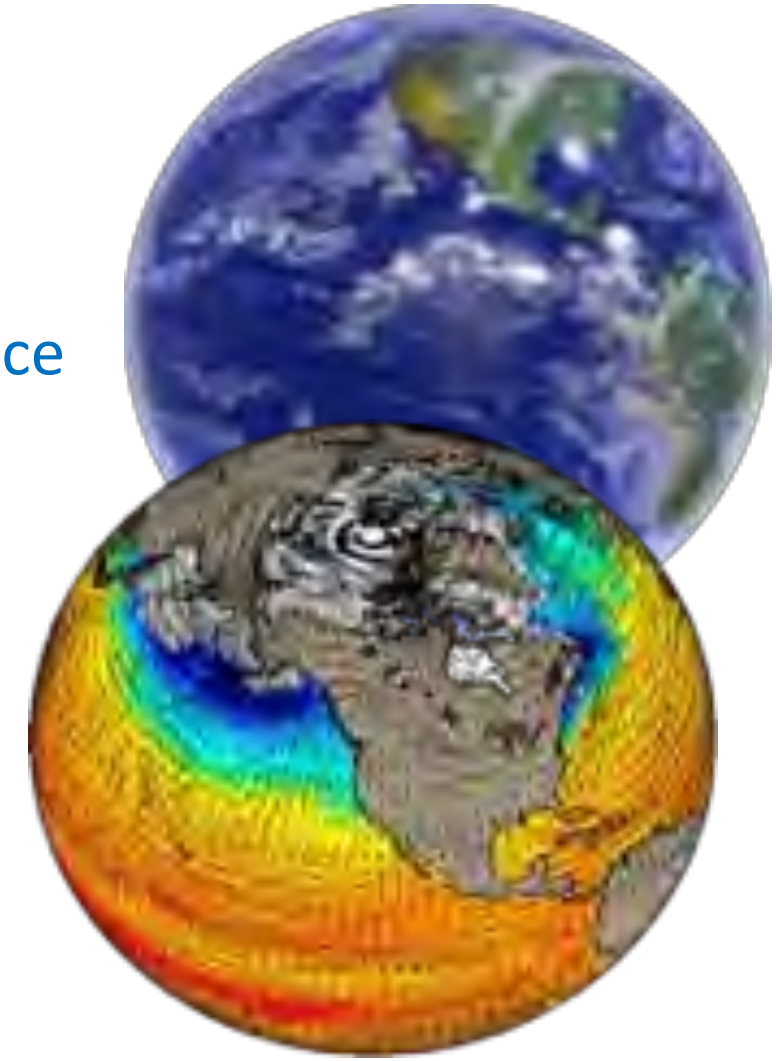
Datasets

Model

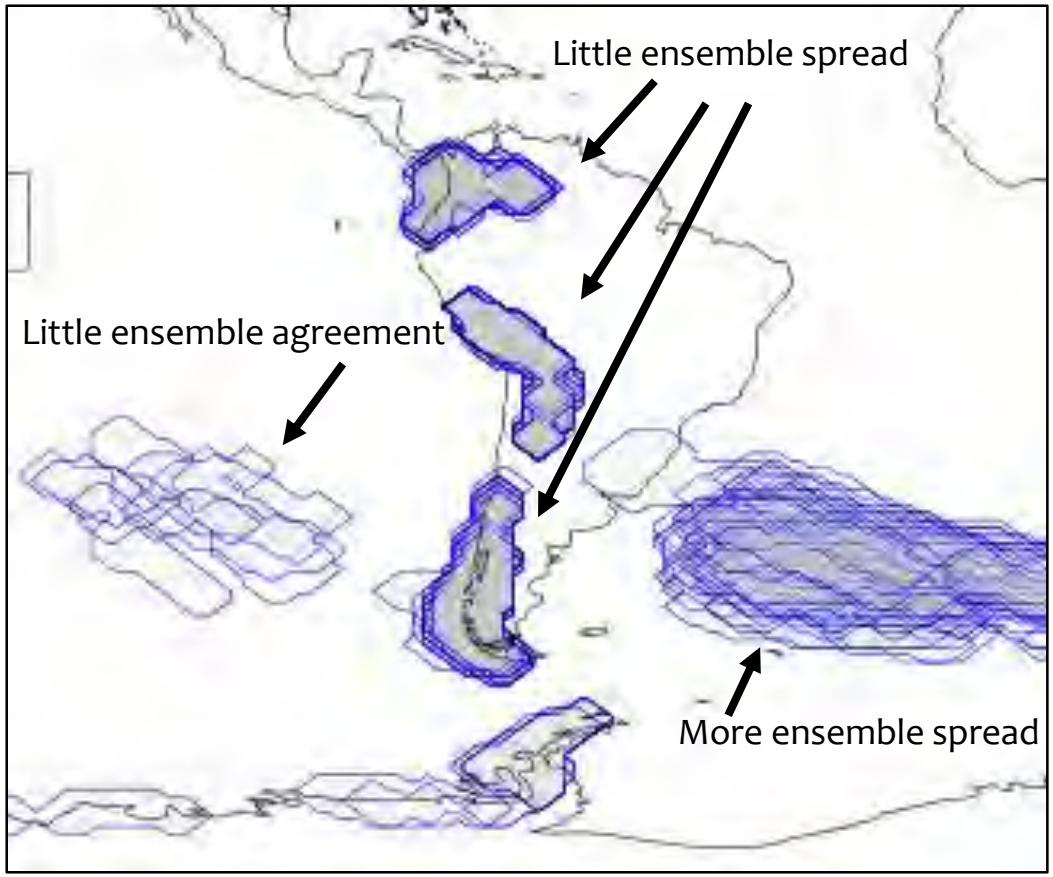
- Global, ~1 degree, 32 members
- Precipitation, temperature, sea ice

Observations, regridded to CESM

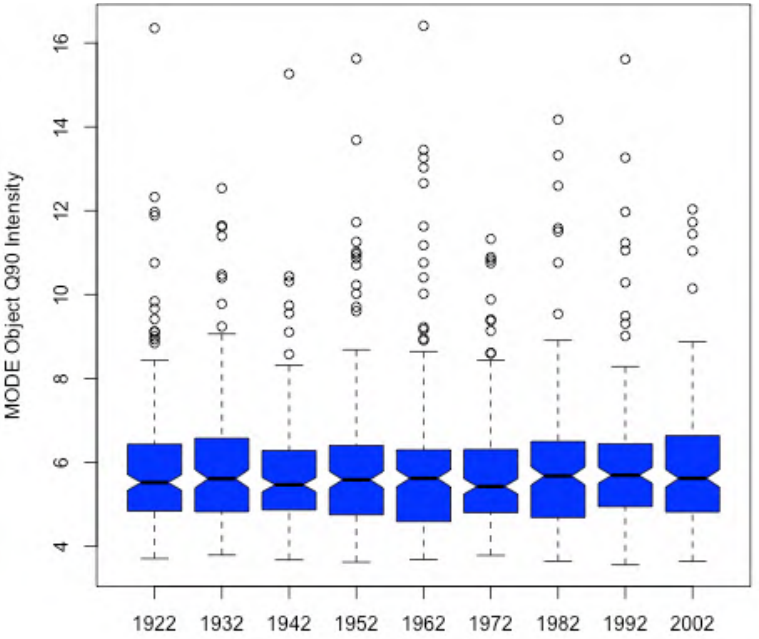
- CHRPS precipitation (Africa)
- Global precipitation climatology project (GPCP)
- Princeton temperature (Sheffield et. al. 2006)
- Sea ice satellite (DMSP-F8 SSM/)



CESM-LE MODE Objects



January Intense Precipitation (90th percentile)



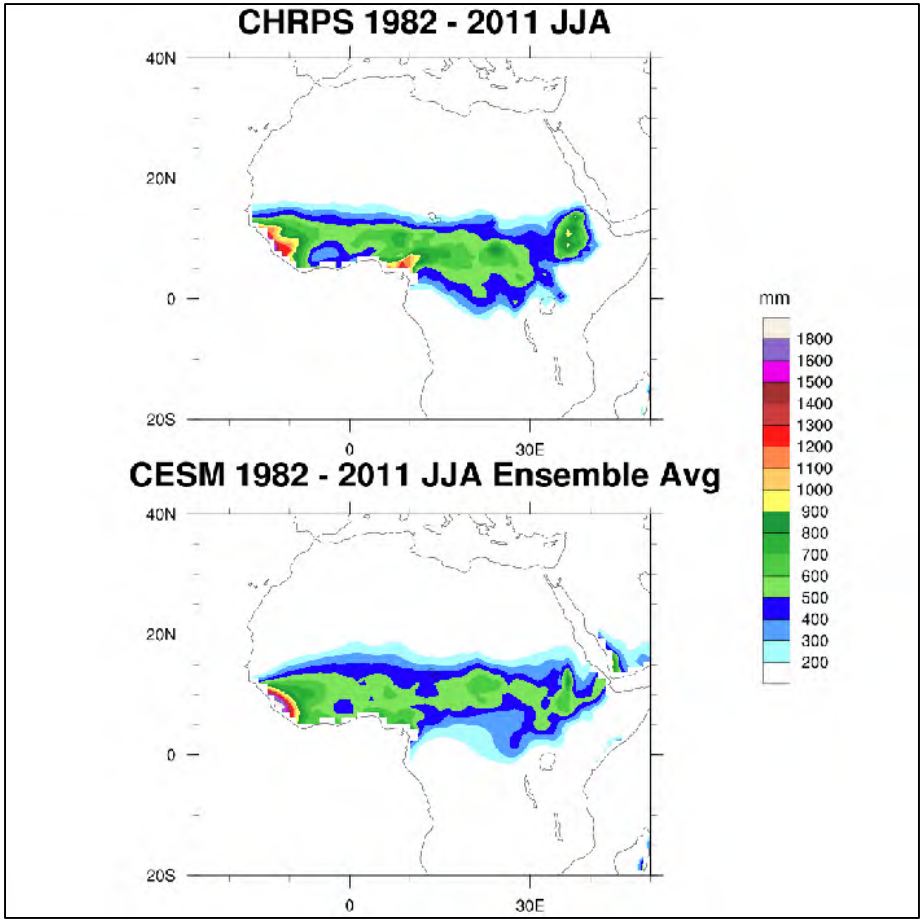
Precipitation objects across 32 members

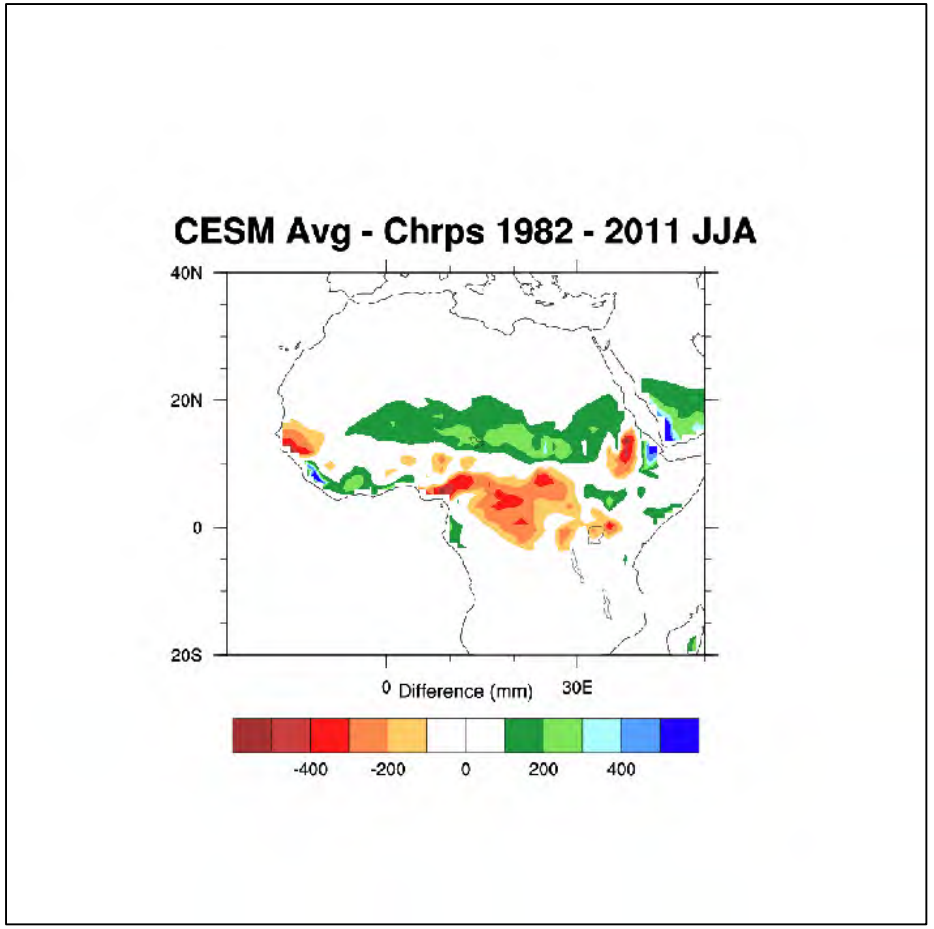
Spatial information on ensemble

Some years have greater spread

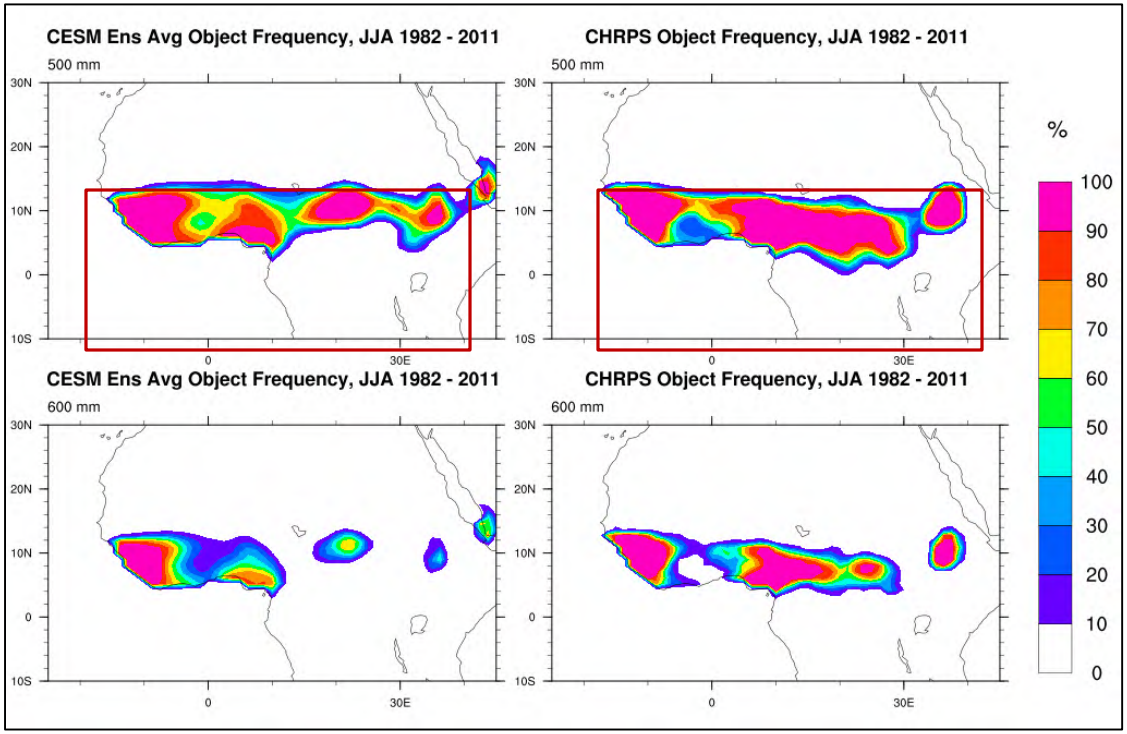
Predictability of extremes

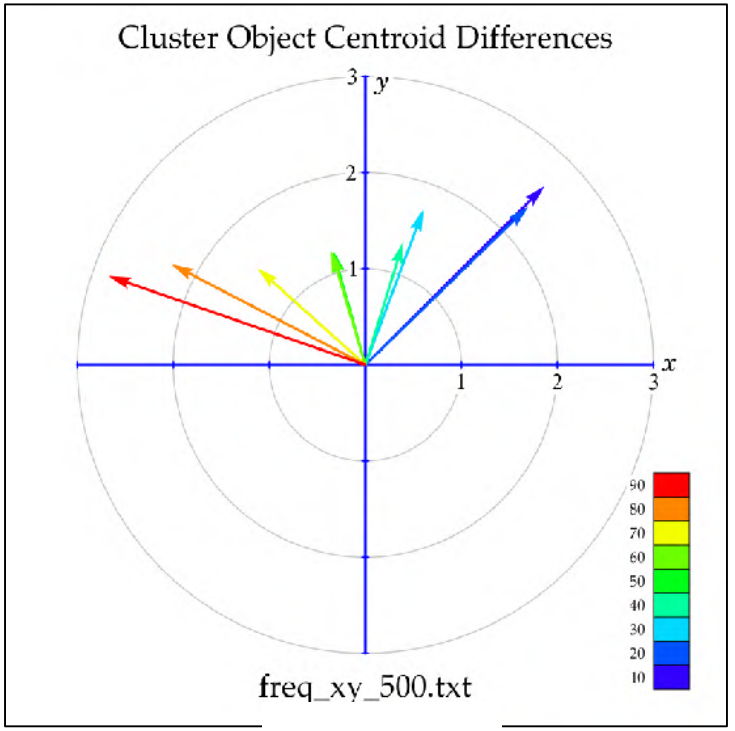
JJA ITCZ Precipitation





JJA ITCZ Precipitation

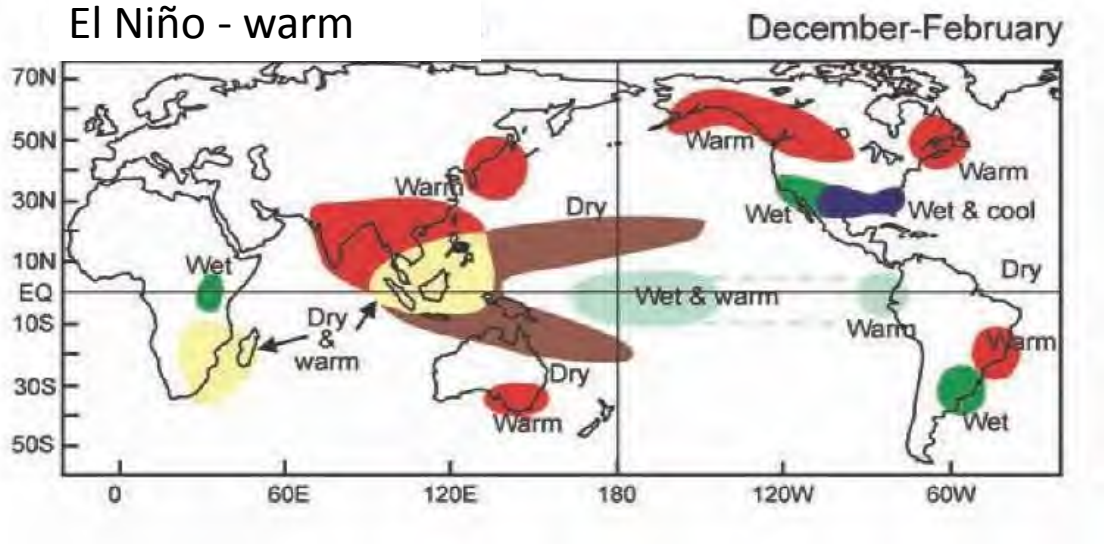




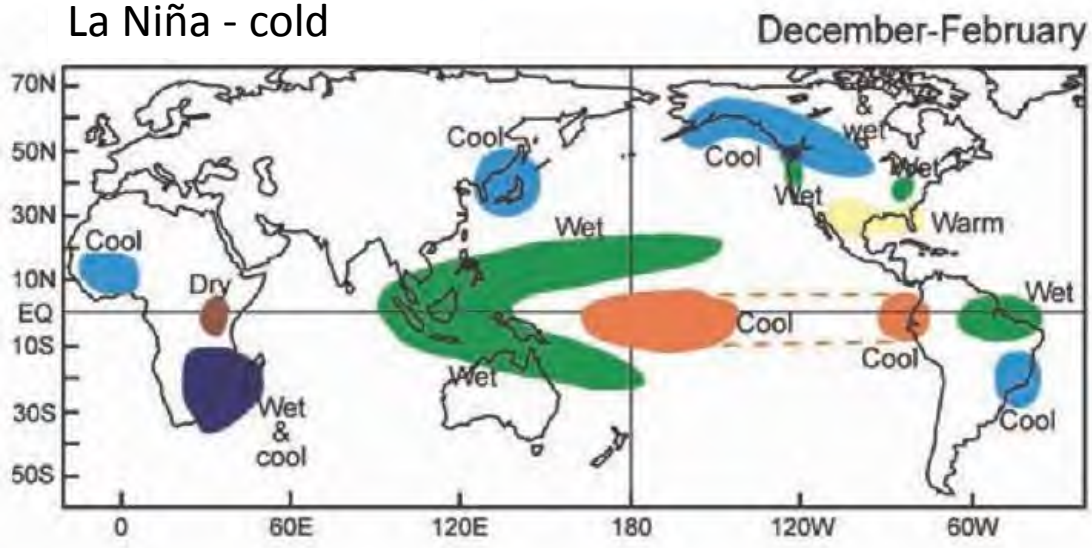
Frequency	Area Ratio	Intersection Area	Symmetric Diff	Centroid Diff
10%	2.64	313	209	2.64
30%	0.93	236	207	1.73
50%	0.93	186	191	1.24
70%	0.86	113	188	1.51
90%	0.70	48	157	2.83

ENSO Variability and Teleconnections

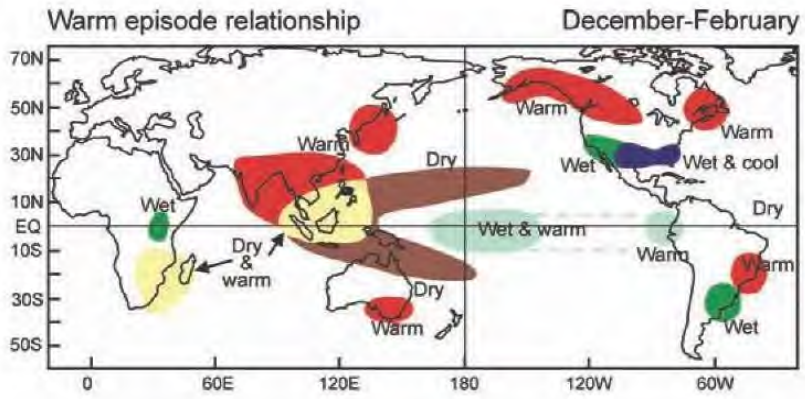
El Niño - warm



La Niña - cold



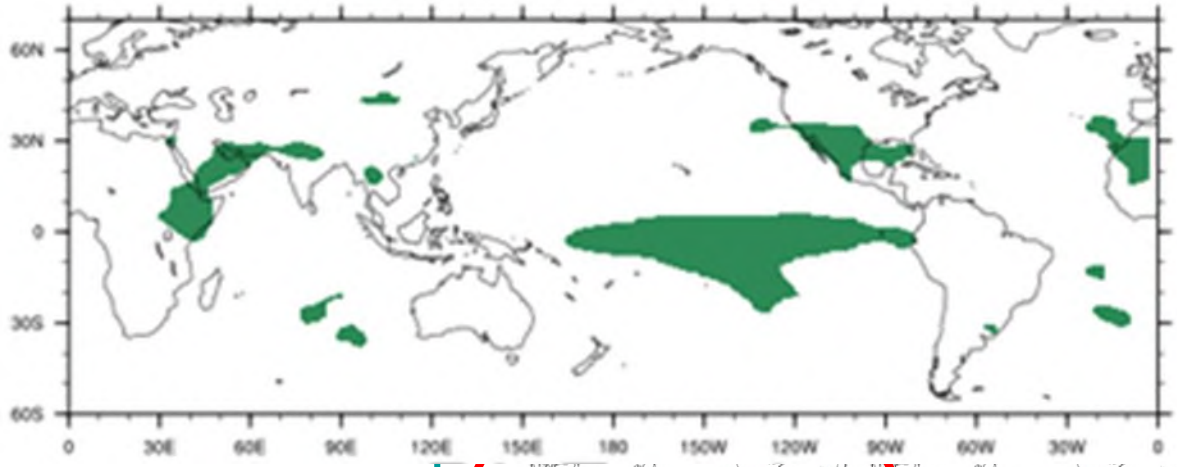
- Can we replicate with model and observations?
- How well do they compare?
- Temperature and precipitation anomalies (1979 – 2015)

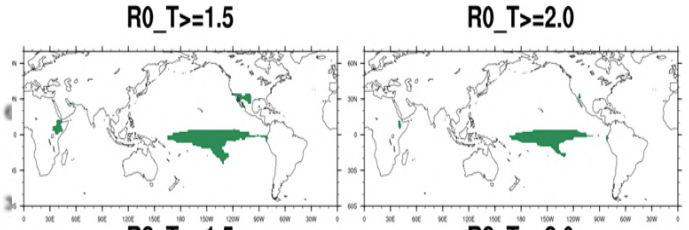


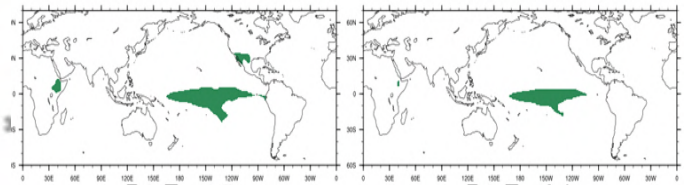
Positive (wet) EN precip anomalies (GPCP)

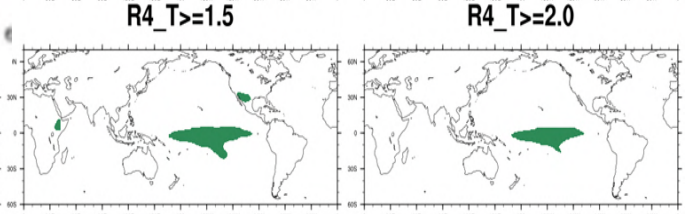
Incr Thresh →

R4_T>=1.0

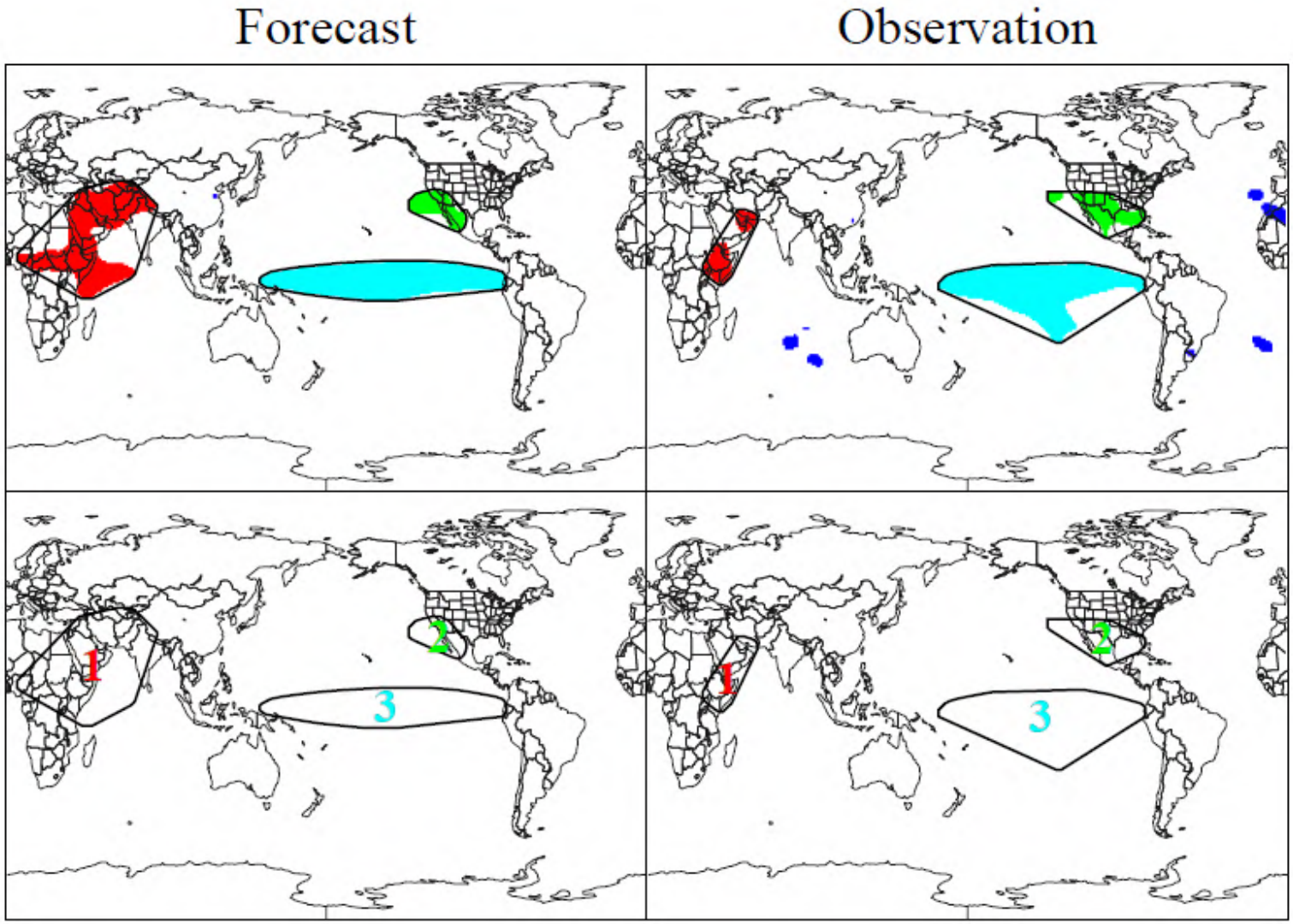




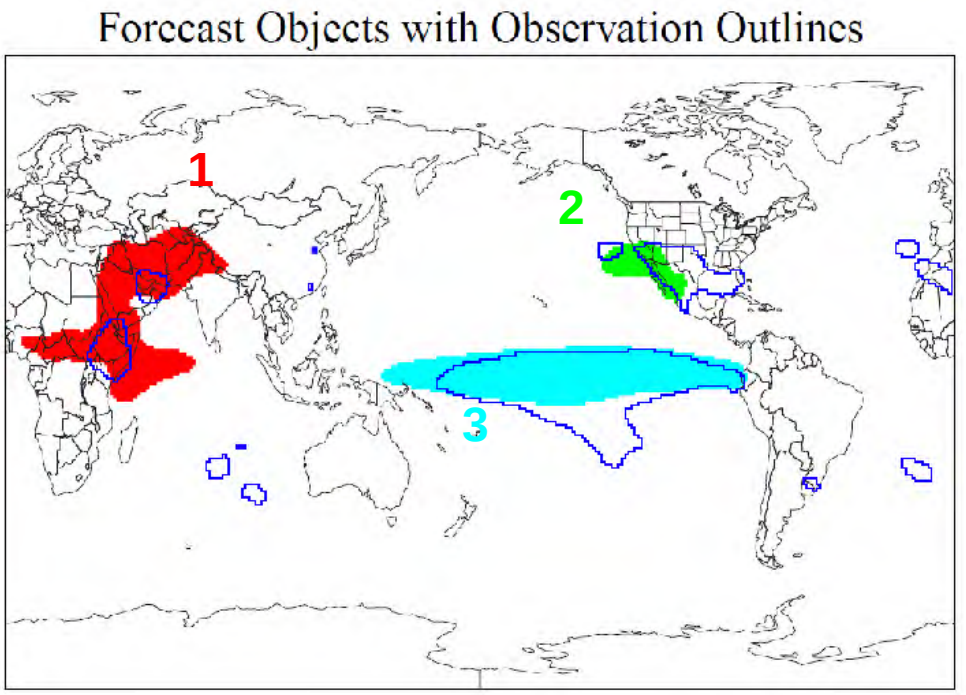




El Nino: CESM vs. GCPC
Wet anomalies

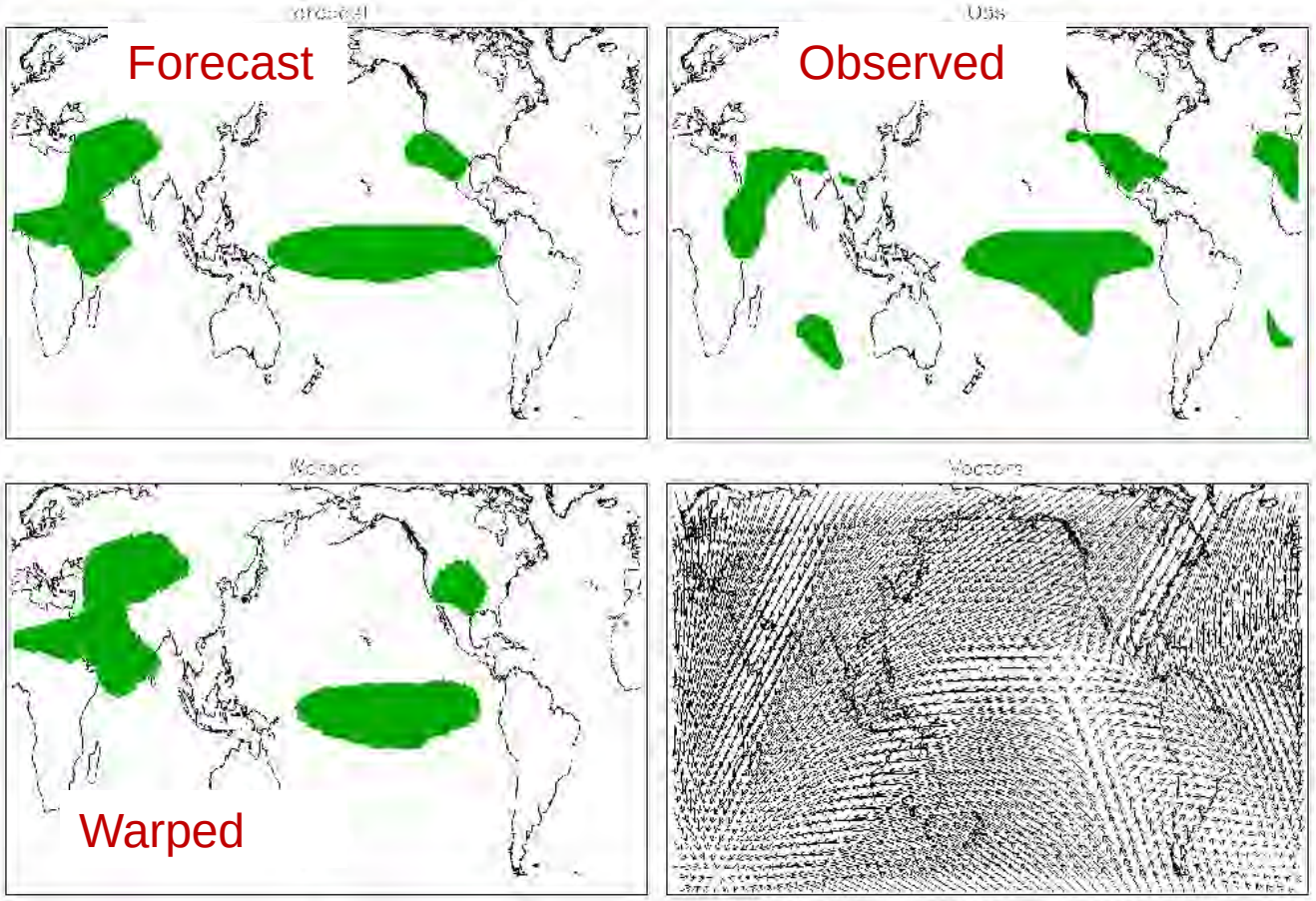


Object Comparisons (EN Wet anomalies)



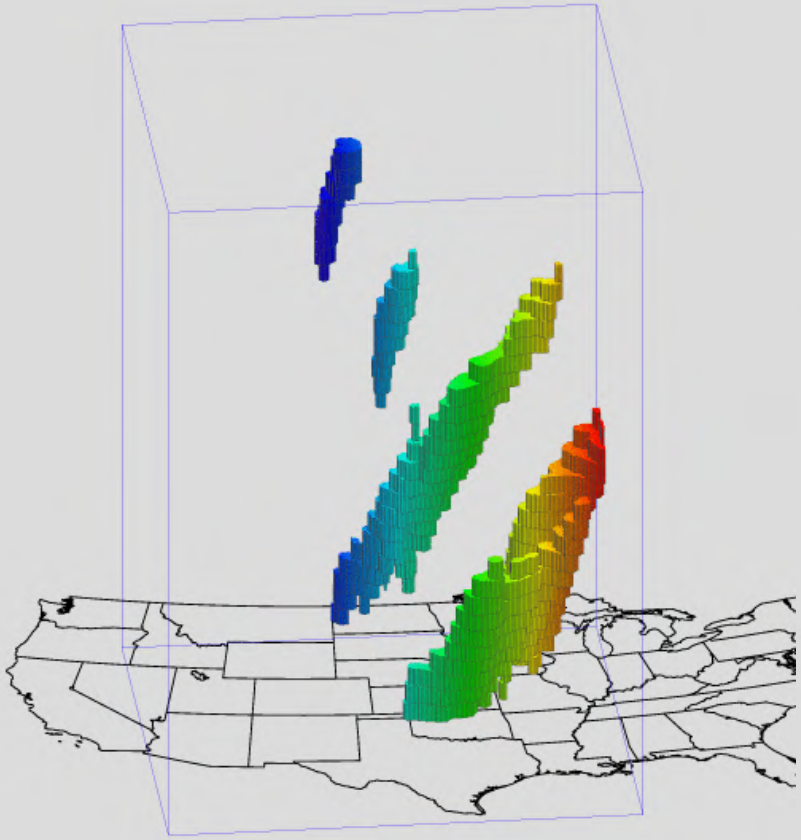
Attribute	Cluster 1		Cluster 2		Cluster 3	
	Fcst	Obs	Fcst	Obs	Fcst	Obs
Area	1269	237	242	333	1405	1498
Median intensity	1.6	1.5	1.2	1.4	3.0	2.2
0.90th intensity	3.0	2.0	1.4	1.9	5.0	4.9
Area ratio (F/O)	5.4		0.73		0.94	
Centroid difference	6.3		10.1		7.7	

Image warping: EN wet anomalies



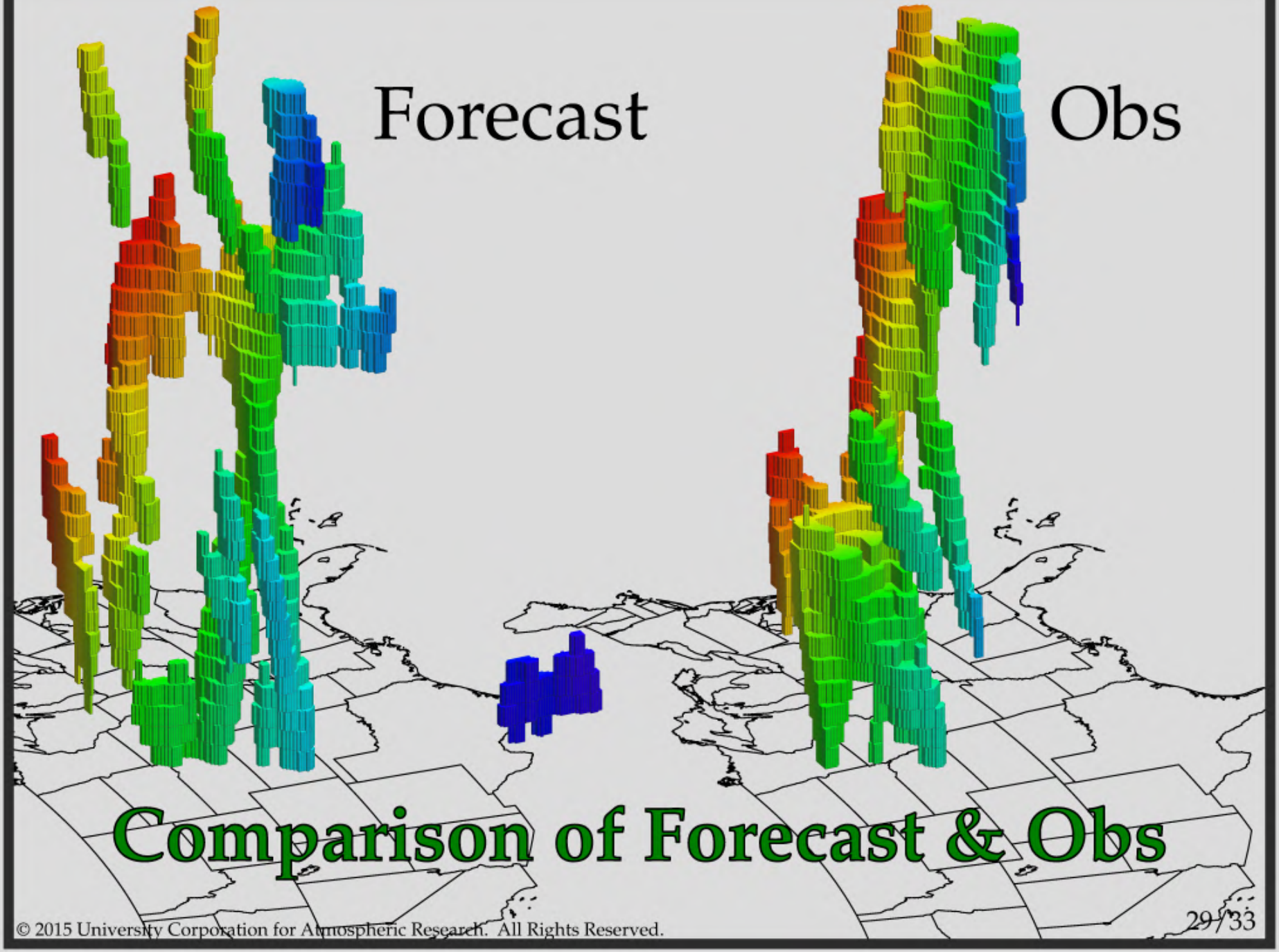
RMSE reduction from warping: 42%

3D Objects



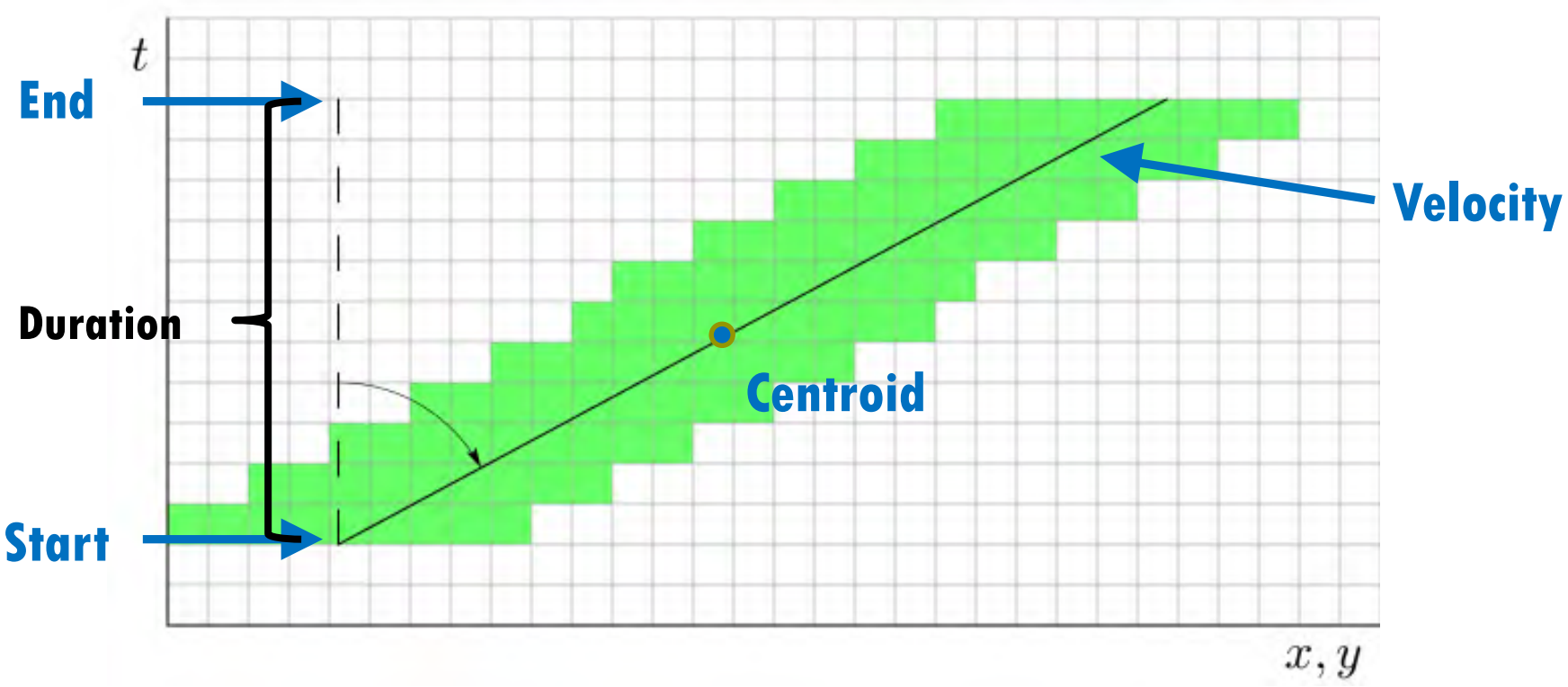
June 13, 2002
IHOP Precip Data

Vertical Dimension
is Time

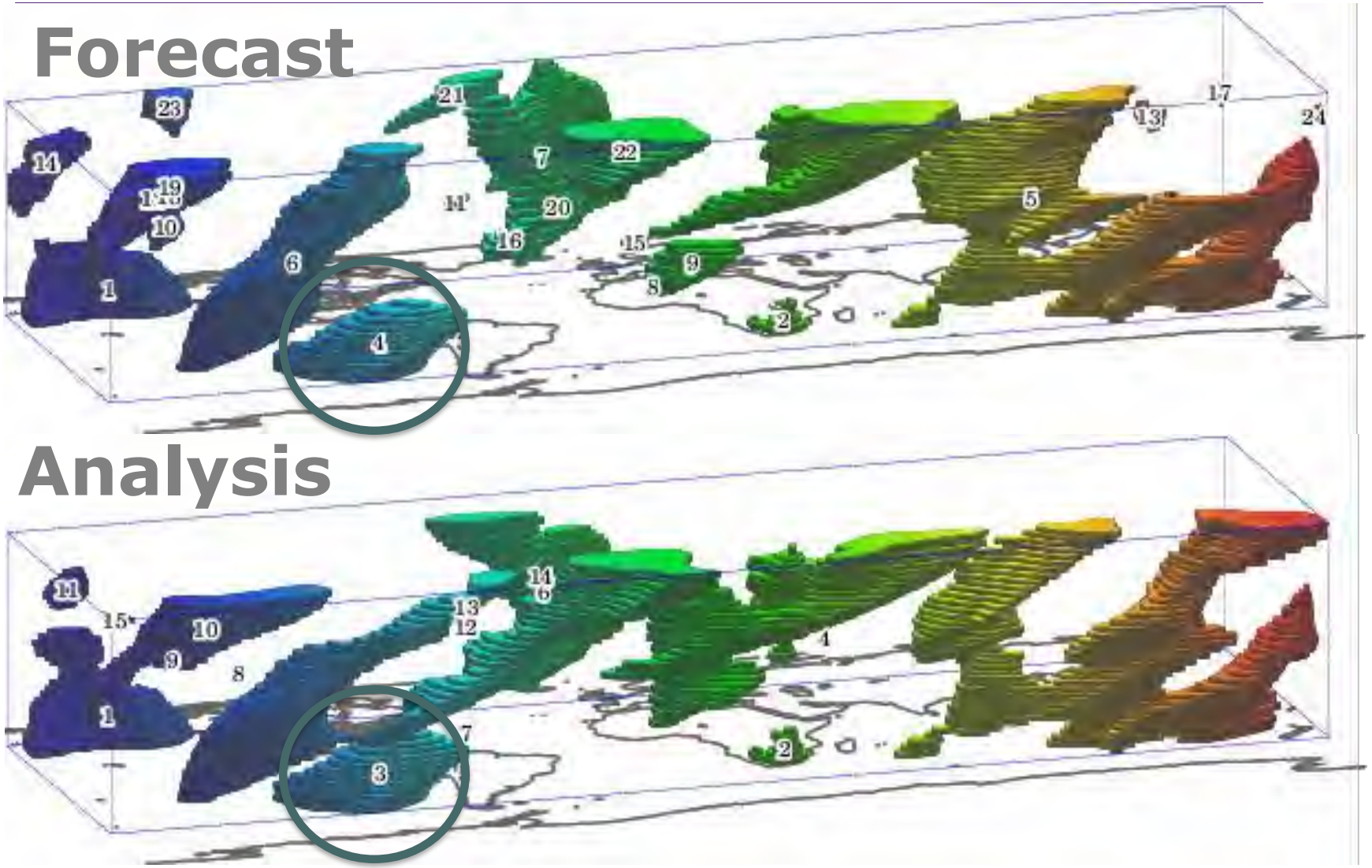


Comparison of Forecast & Obs

Attributes (Think of object as 2D slice)

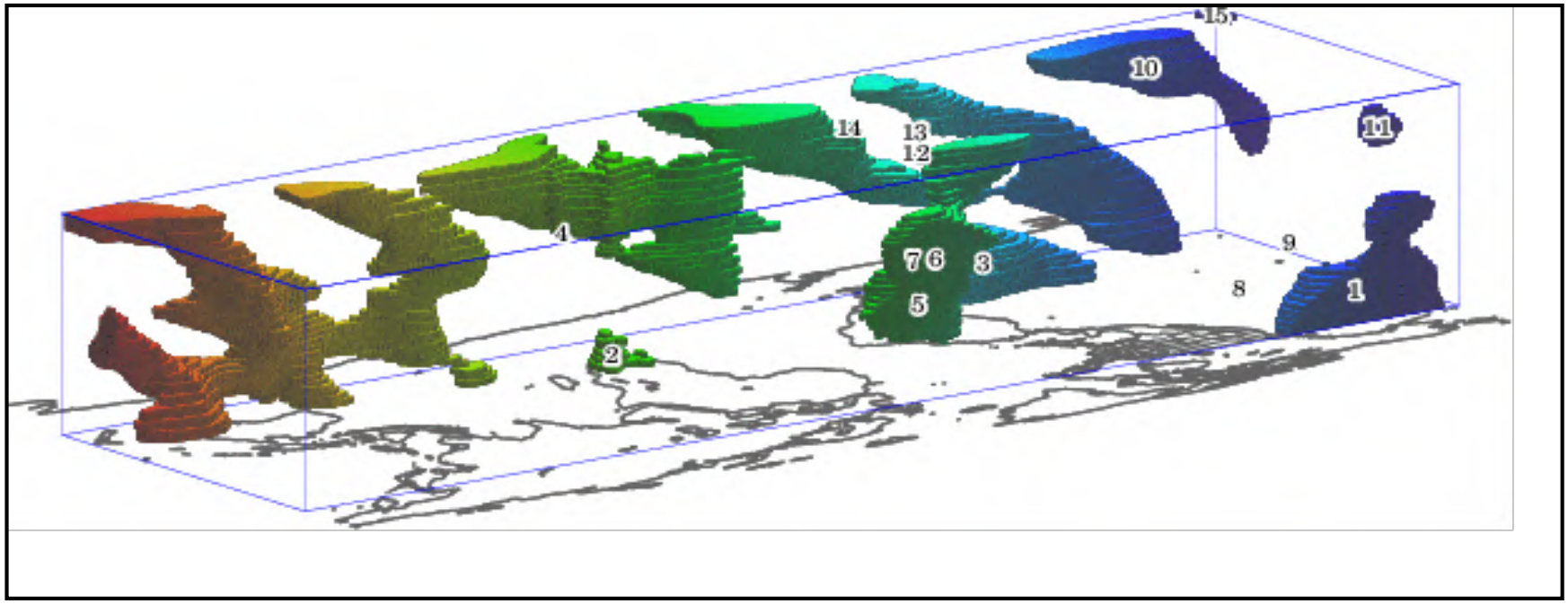


MODE Time-Domain



MODE Time-Domain: High pressure objects (from GFS)

f000 – f240	Max Inten	Volume	Centroid (x,y,t)	Velocity
Fcst Object 4	103927	111493	336, 57, 4.19	2.85
Analysis Object 3	103914	113692	335, 59, 4.27	2.79



Courtesy John Halley Gotway

Summary

- Methods exist for a meaningful evaluation of forecasts of extremes
  - *New methods provide meaningful information that is more appropriate than traditional approaches (based on extreme value theory)*
  - But… still an area of research
- Spatial methods hold much promise for evaluation of climate, S2S and seasonal predictions
  - *Also still an area of research BUT many methods are available to use*

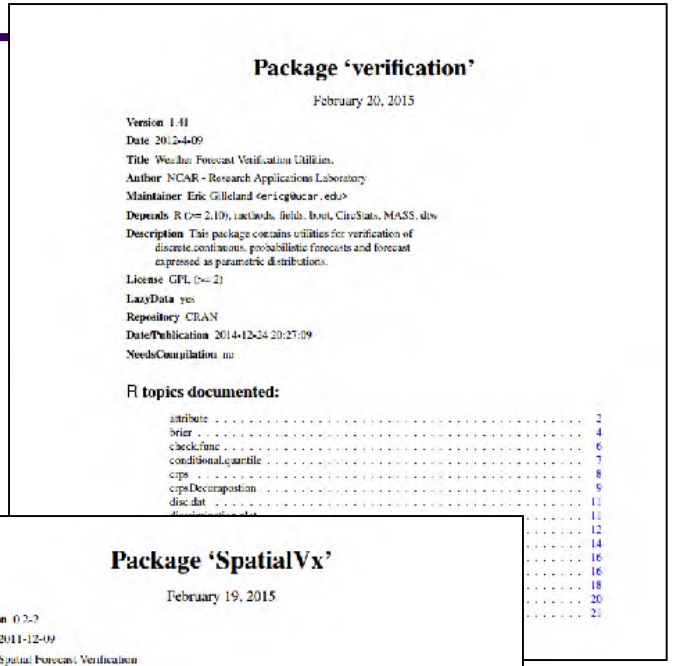
Resources

- Tools
  - MET package (https://dtcenter.org/met/users/)
    - Includes neighborhood methods, MODE, Intensity Scale, SEDI etc.
  - R verification packages (see next slide)
- Literature on spatial methods:

  https://ral.ucar.edu/projects/icp/

R verification libraries

- R is available at http://www.r-project.org/
- Maintained and supported by Eric Gilleland (NCAR)



Package 'verification'

February 20, 2015

Version 1.41
Date 2012-4-09
Title Weather Forecast Verification Utilities.
Author NCAR - Research Applications Laboratory
Maintainer Eric Gilleland <ericg@ucar.edu>
Depends R (>= 2.10), methods, fields, boot, CircStats, MASS, dtw
Description This package contains utilities for verification of discrete,continuous, probabilistic forecasts and forecast expressed as parametric distributions.
License GPL (>= 2)
LazyData yes
Repository CRAN
Date/Publication 2014-12-24 20:27:09
NeedsCompilation no

R topics documented:

Package 'SpatialVx'

February 19, 2015

Version 0.2-2
Date 2011-12-09
Title Spatial Forecast Verification
Author Eric Gilleland <EricG@ucar.edu>
Maintainer Eric Gilleland <EricG@ucar.edu>
Depends R (>= 2.10.0), spatstat (>= 1.37-0), fields (>= 6.8), smoothie, sna(), fortran3EM
Imports distillery, maps, boot, CircStats, fastcluster, waveslim
Suggests shapes
Description Functions to perform spatial forecast verification
License GPL (>= 2)
URL http://www.ral.ucar.edu/projects/icp
BugReports http://www.ral.ucar.edu/projects/icp/spatialvx
NeedsCompilation no
Repository CRAN
Date/Publication 2014-12-24 01:45:06

R topics documented: