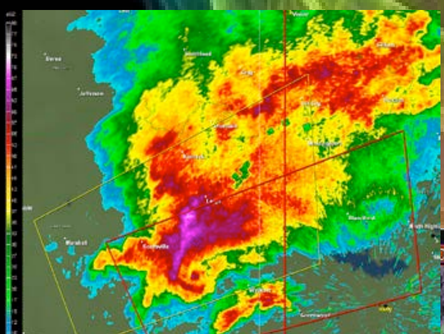
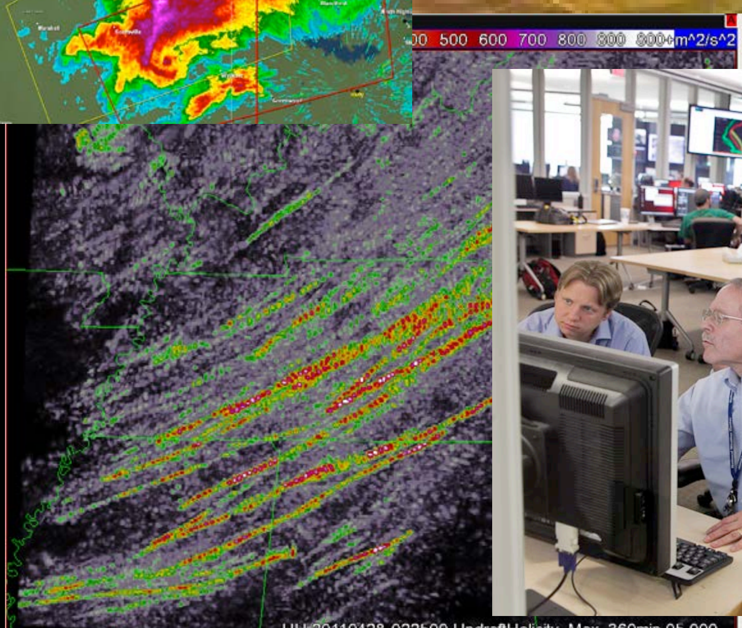
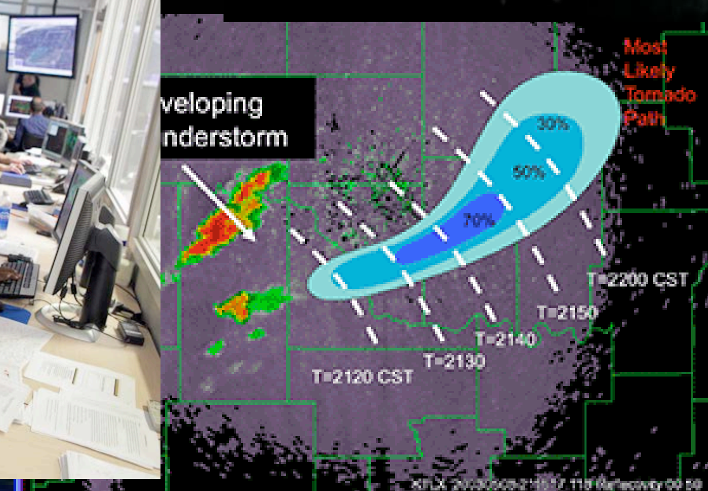
Severe Storm Forecast Verification

Adam J. Clark
Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma and NOAA/National Severe Storms Laboratory

October 23, 2014 – International Cooperative for Aerosol Prediction (ICAP) Workshop







Outline – Severe Storms Verification

- Verification of SPC outlooks – brief history
  - Hitchens and Brooks papers…
    - Hitchens, N. M., and H. E. Brooks, 2012: Evaluation of the Storm Prediction Center's Day 1 Convective Outlooks. *Wea. Forecasting*, **27**, 1580-1585.
    - Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525-534.
    - Hitchens, N. M., and H. E. Brooks, 2014: Evaluation of the Storm Prediction Center's convective outlooks from day 3 through day 1. *Wea. Forecasting*, **29**, 1134-1142.
- Verification of NWP forecasts of severe storms
  - NOAA/HWT Spring Forecasting Experiments
  - DTC Visitor Program projects

Verification of Storm Prediction Center convective outlooks

- Convective Outlooks (COs) – Primary means by which severe weather risk over US communicated to the public.
  - Day 1 COs issued at 0600 UTC daily and cover the 1200 to 1200 UTC period.
  - "Severe" defined as: tornado, hail ≥ 1 in, and/or wind speed ≥ 50 knots within 25 miles of a point
  - "Significant severe" defined as: tornado EF2+, hail ≥ 2 in, and/or wind speed ≥ 65 knots within 25 miles of a point
  - Three different categorical risk levels: slight, moderate, and high – each is associated with a probability. For Day 1, categorical level depends on probability associated with specific severe weather type. For Day 2, categorical level depends on probability for total severe.

Day 1 Probability to Categorical Outlook Conversion

(SIGNIFICANT SEVERE area needed where denoted by hatching - otherwise default to next lower category)

Outlook Probability	TORN	WIND	HAIL
2%	SEE TEXT	NOT USED	NOT USED
5%	SLGT	SEE TEXT	SEE TEXT
10%	SLGT	NOT USED	NOT USED
15%	MDT	SLGT	SLGT
30%	HIGH	SLGT	SLGT
45%	HIGH	MDT	MDT
60%	HIGH	HIGH	MDT

Day 2 Probability to Categorical Outlook Conversion

(SIGNIFICANT SEVERE area needed where denoted by hatching - otherwise default to next lower category)

Outlook Probability	Combined TORN, WIND, and HAIL
5%	SEE TEXT
15%	SLGT
30%	SLGT
45%	MDT
60%	HIGH

Verification of SPC COs (cont)

- "Easy" to verify!
  - Forecasts cover long time periods and large scales.
  - Database of reports matching severe weather definition extending to 1950.
    - Huge caveat – database heavily impacted by non-meteorological influences (changes in reporting practices, better communication, storm-chasers, population density, etc).
  - "Traditional" or contingency table based metrics can be applied.
  - Hitchens and Brooks wanted to answer: How accurate are COs? Has skill changed over time? Does skill improve with decreasing lead time?

Description of Data

- Convective outlooks – available since 1973
  - Slight risk areas plotted on 80 km × 80 km grid
    - Approximately equivalent to SPC's "25 miles of a point"
    - Other grid sizes used as well to test impact on skill
- Storm reports
  - Plotted on same grid as convective outlooks
  - Considered dichotomous events
    - Grid box either "yes" or "no" regardless of report count

Verification Measures (comprise Roebber et. al 2009 performance diagram)

		Observed	
		Yes	No
Forecast	Yes	a	b
	No	c	d

→

Hit	False Alarm
Miss	---

POD = a / (a + c) (fraction of events correctly forecast)

FOH = a / (a + b) (fraction of forecasts that were correct)

CSI = a / (a + b + c) (fraction of observed and/or forecast events correctly predicted)

Bias = (a + b) / (a + c) (ratio of forecast to observed events)

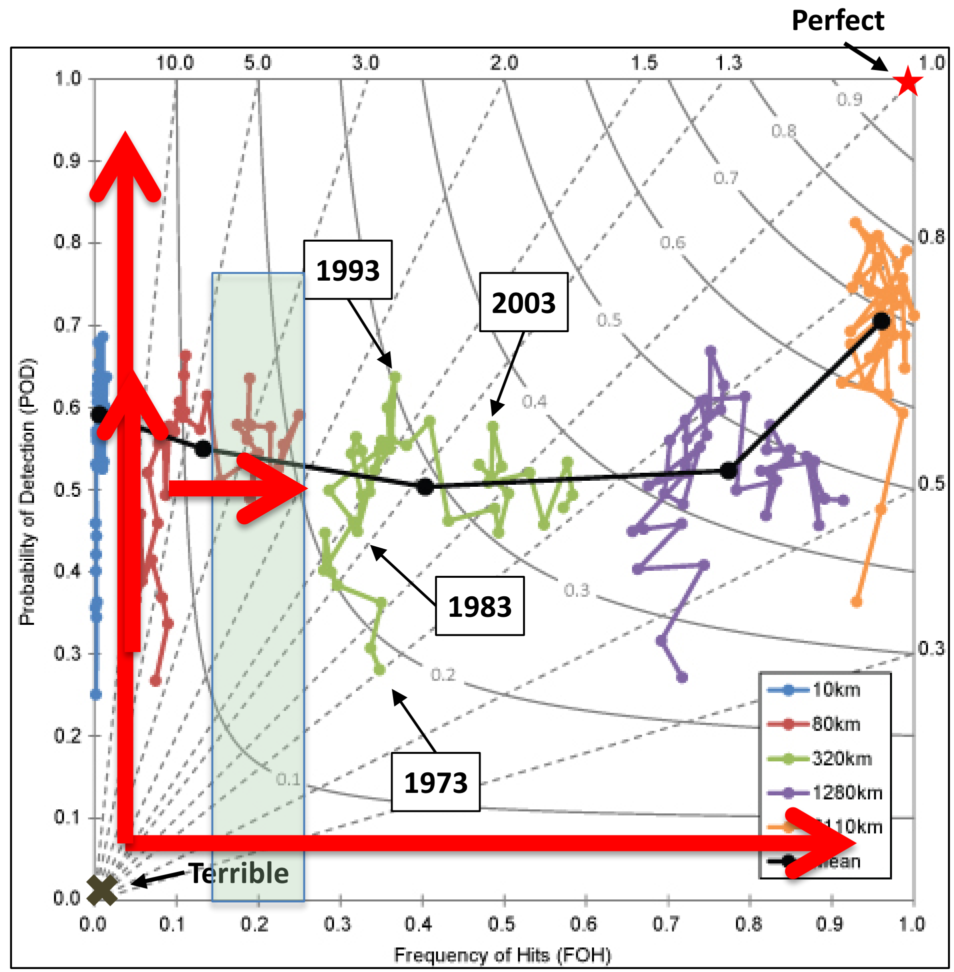
Because CSI and bias can be expressed in terms of POD and FOH, all four measures can be represented on same diagram.

Performance Diagram

- Different colors for different grid sizes. For each color, points are for different years. Black line is mean at each scale.
- FOH at 80-km 0.15 to 0.25, which matches probability range for slight risk. SPC forecasters are “reliable”!

- **Trends with grid size:**
  - FOH increases with coarser grid (fewer forecast events, but greater percentage of forecasts are hits).
  - POD stays nearly the same with coarser grid (percentage of observed events correctly forecast remains constant).

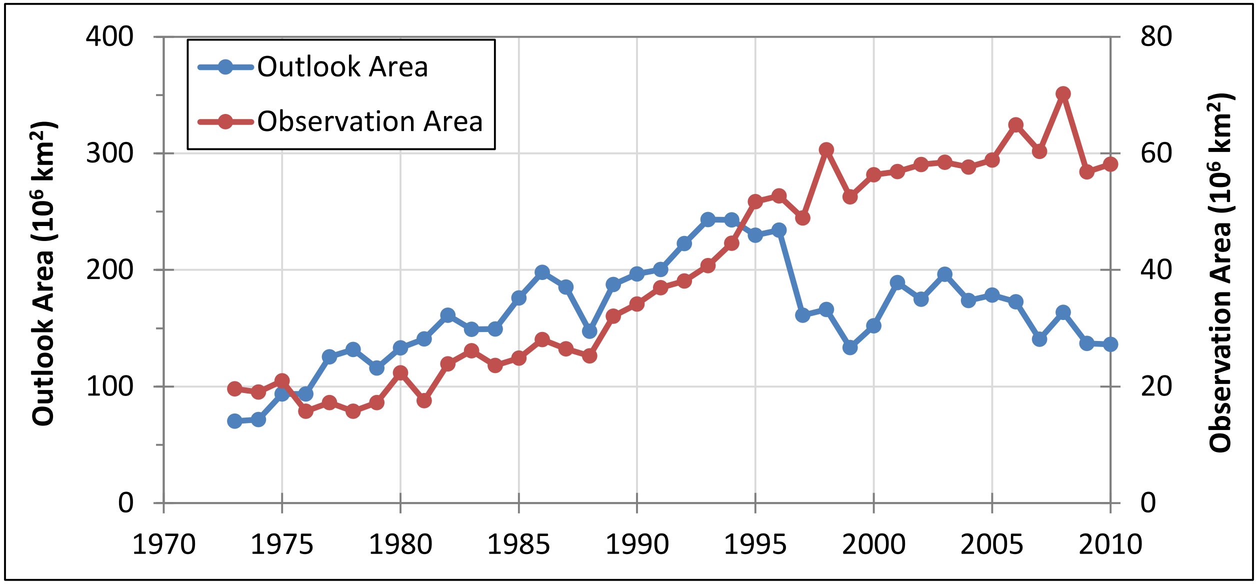
- **Trends with time**
  - Large increase in POD over first 20 years, while FOH increase less dramatic.
  - FOH increases most in remaining years.
  - For 80-km, bias remained fairly steady, but decreased during last 20 years.
  - Continual increase in CSI.
  - What is going on? Clues by looking at trends in areal size of outlooks report coverage (focus on 80 km grid).



CSI = curved lines; Bias = dashed lines (Roebber 2009)

Credit: Nathan Hitchens

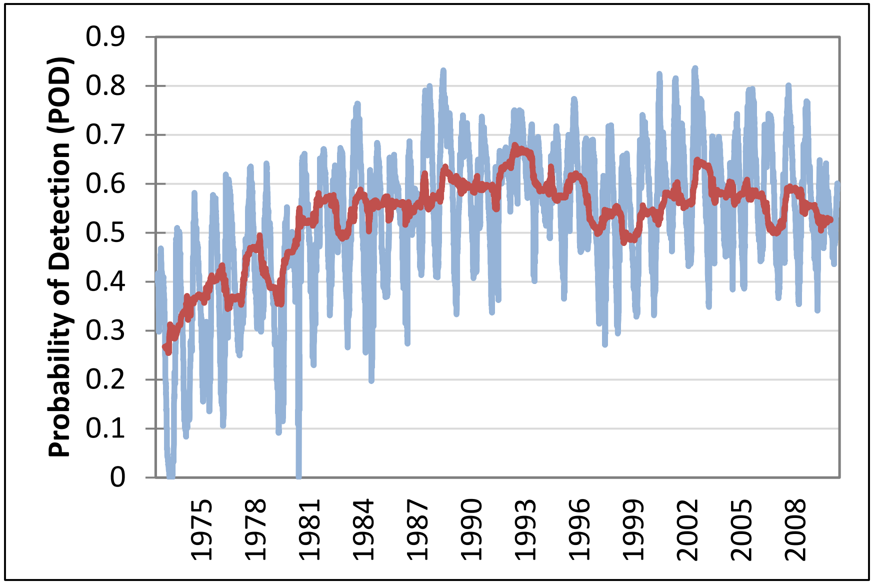
Areal Size – Outlooks & Reports

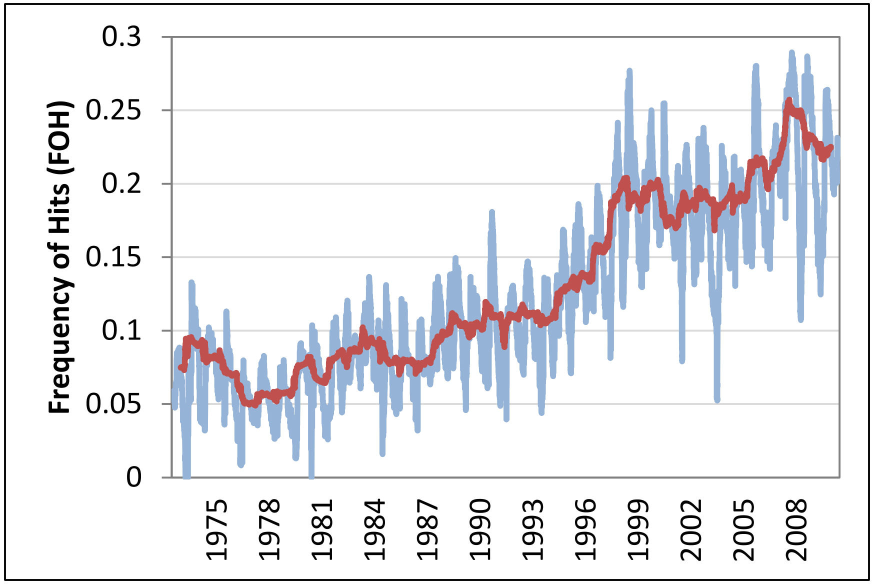


- Observation area increases continually (reflects increased reporting) - Largest increase during 1990s.
- Outlook area peaked during mid-1990s - Leveled off since 1999.
  - Why?
  - Change in forecasting philosophy (increased sensitivity to false alarms?), with one factor being organizational restructuring and influx of new forecasters preceding physical relocation of SPC during 1995-97 time period.

Credit: Nathan Hitchens

POD & FOH





- 365-day running means (red lines)
- 91-day running means (blue lines)
- Outlooks capturing larger fraction of reports, then outlooks becoming more precise with larger fraction of correct forecasts over time.

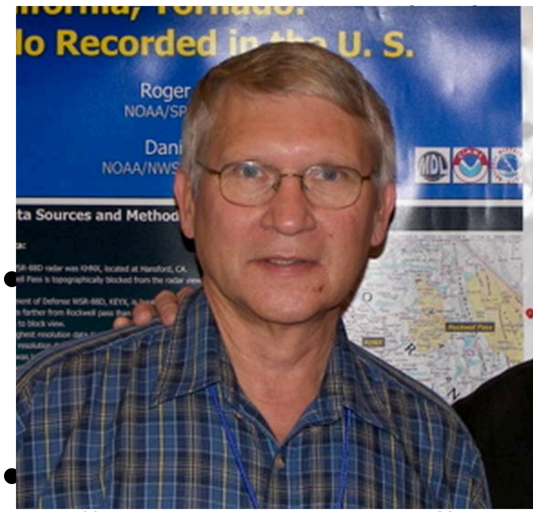
Credit: Nathan Hitchens

Evaluating Skill: "Practically Perfect" (PP) Forecasts

- Skill Should be defined relative to some baseline. For outlooks, what is best baseline?

- Misses/False alarms expected.

- Hitchens and Brooks describe a PP forecast as, "… *a forecast that is* … *t w* … *d m* … *epo* … *d a* … *tea* … *yst*

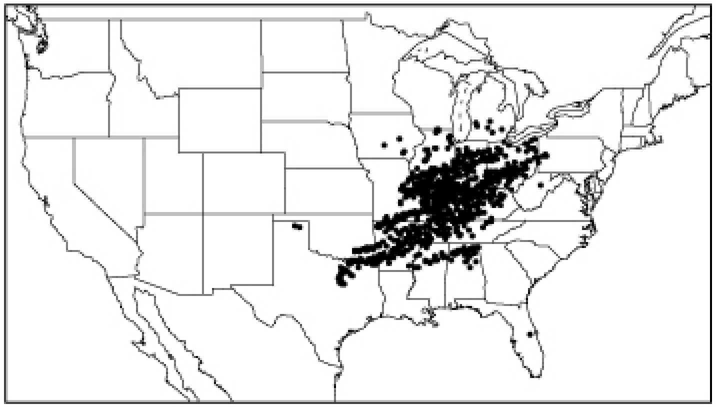



- … ing … me … o d … ern

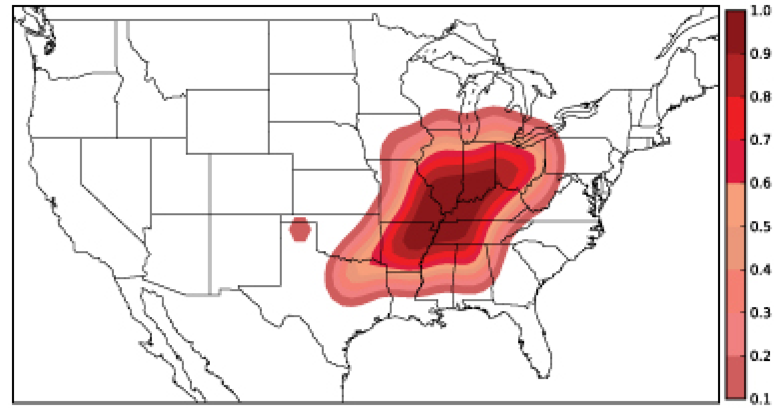
- … act … gh "outbreak days".

Credit: Nathan Hitchens

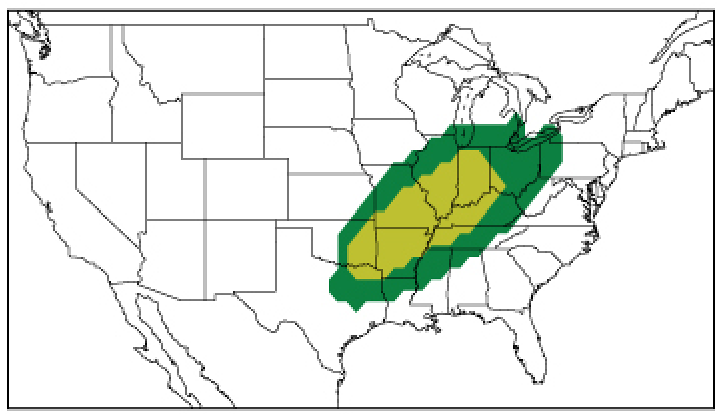
19 April 2011



Storm Reports



Practically Perfect Forecast



0600 UTC Day 1 Convective Outlook

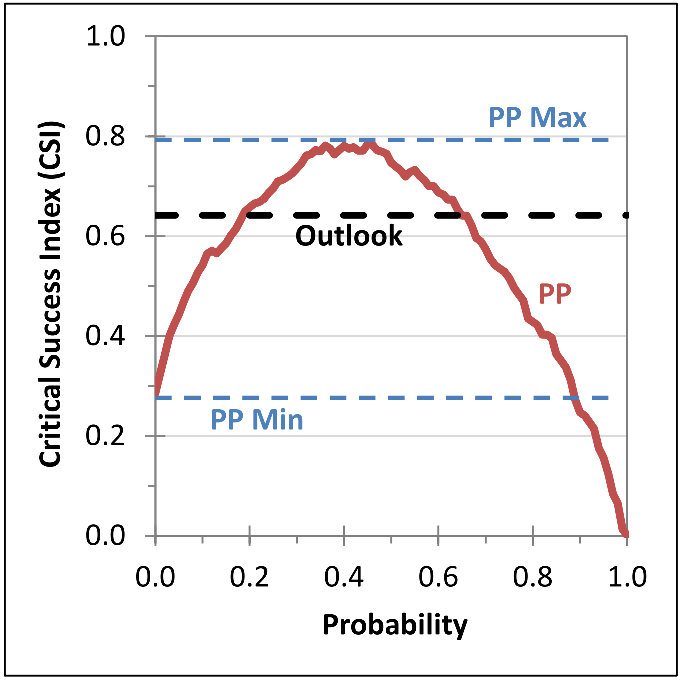
Using characteristics of convective outlooks ideal parameters for PP forecasts chosen.

Credit: Nathan Hitchens

PP as Baseline for Skill

- *Practical* maximum and minimum CSI values for each day can be determined using PP forecasts

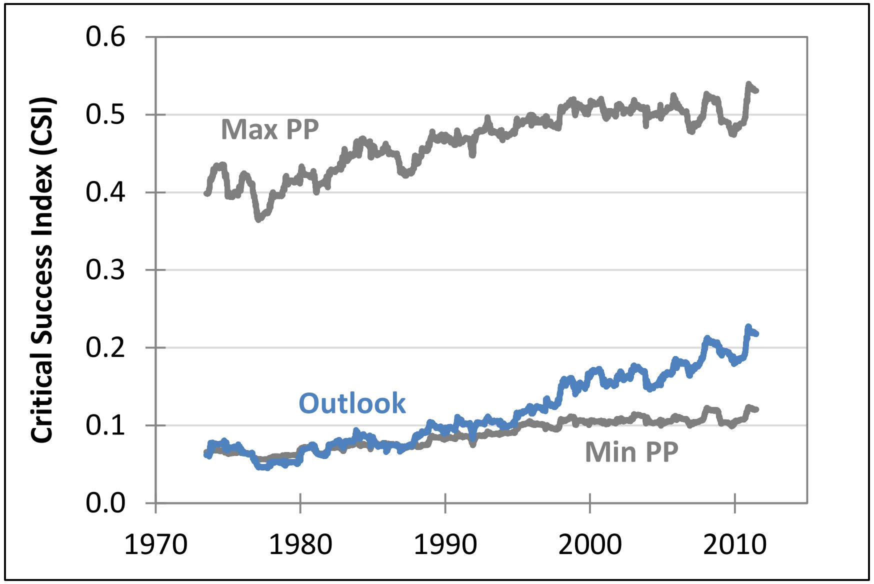
- Example case
  - CSI = 0.64
  - Relative skill = 0.71
    - Max = 0.78; Min = 0.29

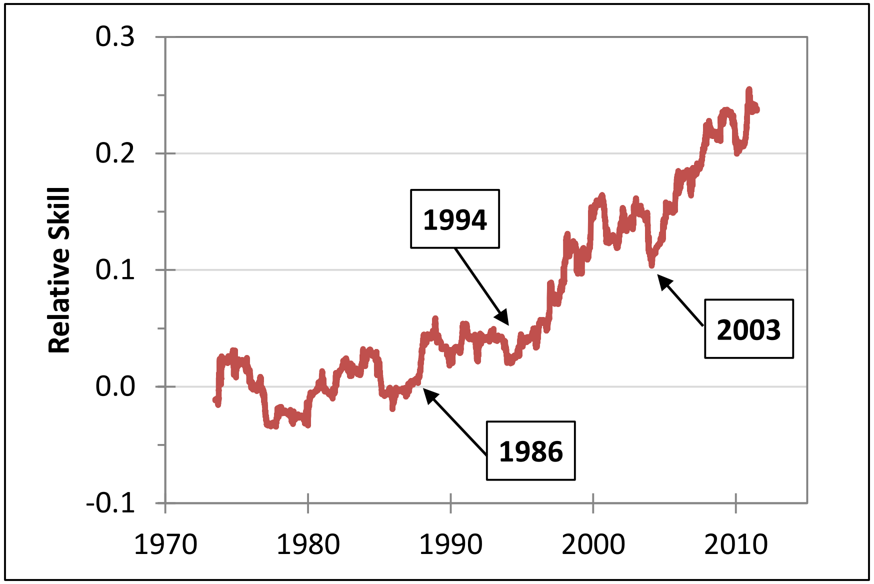


19 April 2011

Credit: Nathan Hitchens

Relative Skill





- 365-day running means
  - Computed by constructing 2 × 2 table that sums all 365 forecasts centered on each day
- Relative skill doesn't really start to increase until the mid-1990s.

Credit: Nathan Hitchens

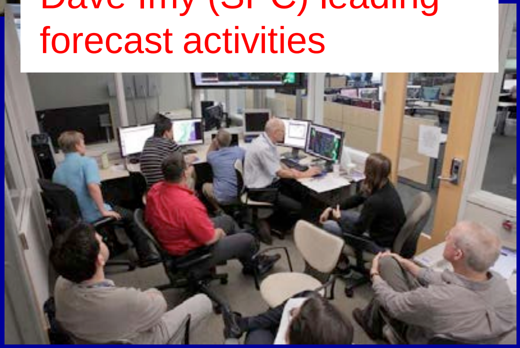
Verification of NWP forecasts of severe storms

- Focus on "convection-allowing" models (CAMs): Grid-spacing ≤ 4-km, coarsest scale you can allow convective overturning to occur on grid-scale and get reasonable results.
- CAMs have been focus of annual NOAA/Hazardous Weather Testbed Spring Forecasting Experiments since 2004.
  - 5-week experiment conducted each spring by SPC/NSSL to evaluate emerging scientific concepts and tools in a simulated operational forecasting environment
  - Primary goals: (1) Accelerate transfer of promising new tools from research to operations (R2O), (2) inspire new initiatives for operationally relevant research (O2R), (3) document performance/sensitivities of CAMs.
  - Clark, A. J., and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55-74.

Scenes from 2013 SFE

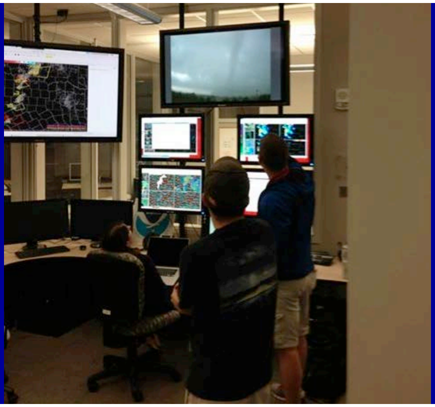
Hand analyses

Dave Imy (SPC) leading forecast activities



Week 2 participants

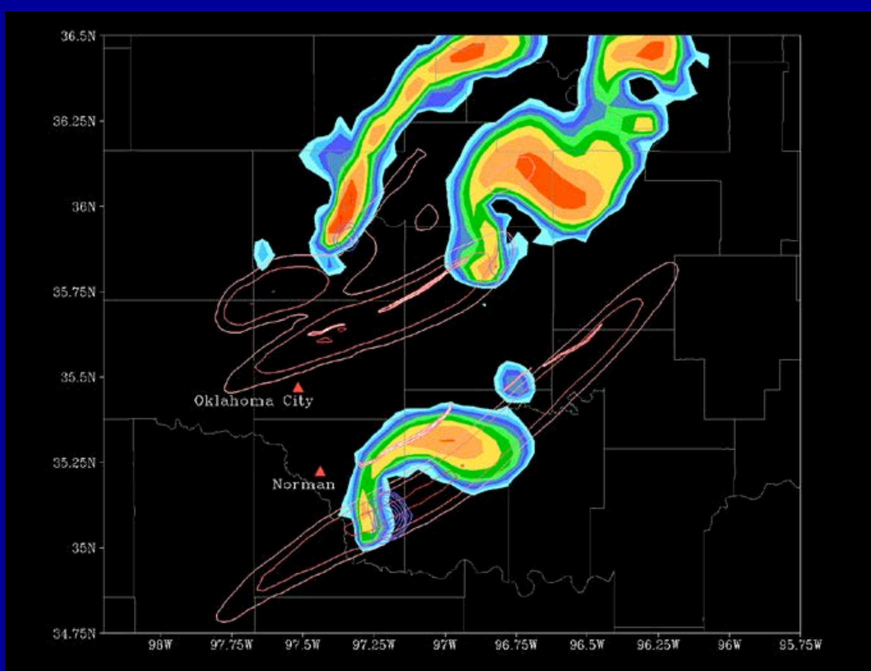
Live stream tornado

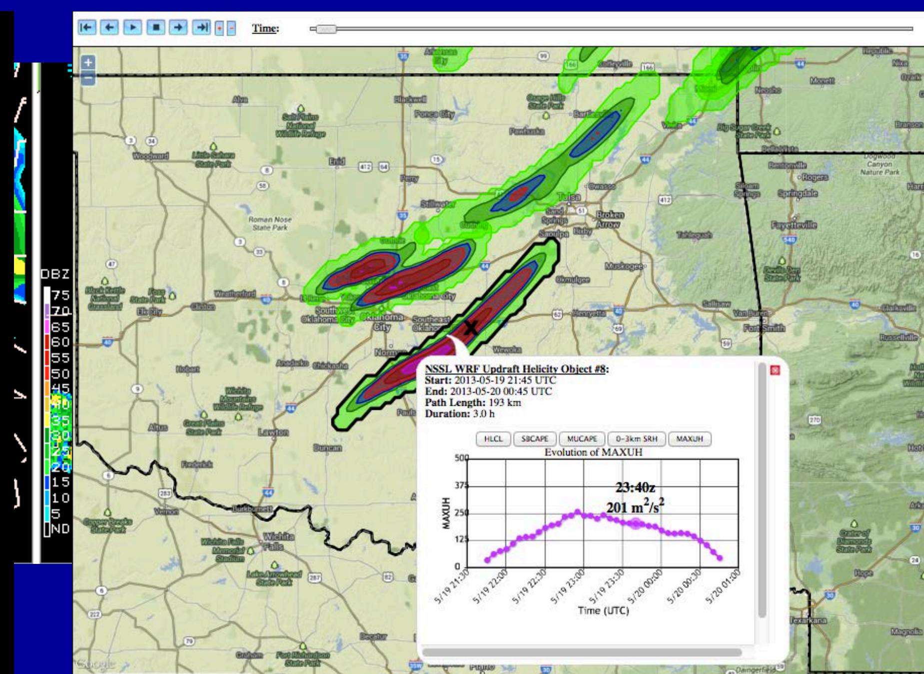


Moore tornado path

Verification of NWP forecasts of severe storms

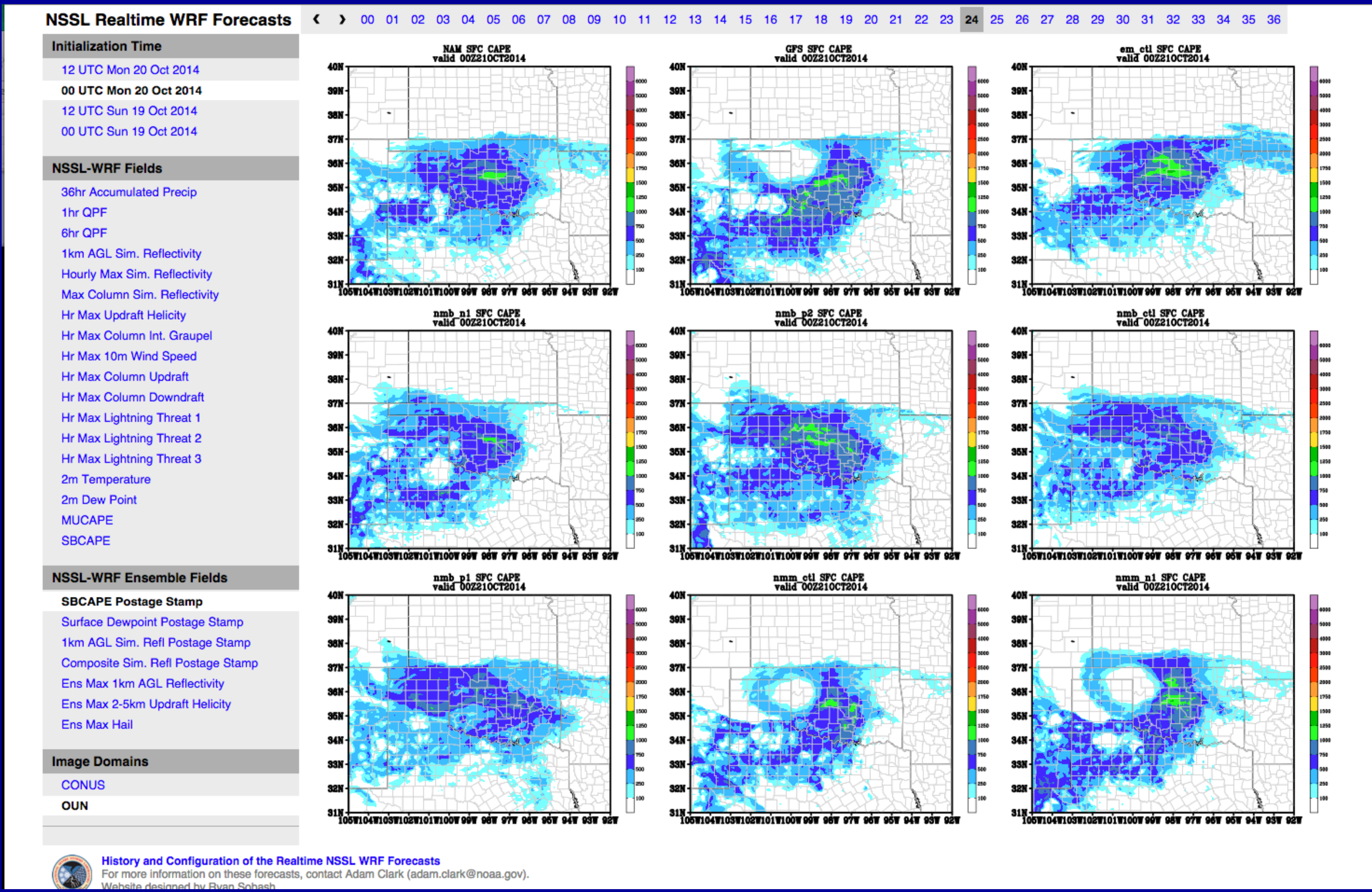
- Findings from SFEs
  - CAMs depict realistic convective scale storm structures
  - Accurately distinguish dominant convective modes
  - At times, provide extraordinarily accurate forecasts of convective system location/timing up to 36 h in advance. Examples…





NSSL-WRF Ensemble -

http://www.nssl.noaa.gov/wrf/newsite



Verification of NWP forecasts of severe storms

- New paradigm needed for CAMs. Rather than only being able to provide info on forecast severe weather environment, CAMs also provide direct info on explicitly simulated storms and related hazards.
  - To fully exploit CAMs requires new and innovative model diagnostics, verification, and visualization strategies.
  - Ensembles are needed to account for the oftentimes very fast error growth at convective scales.

- Many verification challenges!
  - Models still too coarse to directly predict hazards – severe weather "proxies" must be used.
    - Right now, best proxies are UH, Hail, and max 10-m wind.
  - Severe weather observations can be very unreliable at scale of model output.

General issues for severe storms verification

- To extract useful information, need to go beyond traditional metrics (ETS, bias, Brier Score, ROC curves, etc.)
  - Traditional scores can give useful info on larger scale environmental fields, but for short-term forecasts of severe storms, additional methods are needed.
- For severe storms, specific attributes should be verified.
  - storm size, duration, number, timing of CI, intensity of rotation, length of rotation track…
- Ensemble characteristics are important: dispersion, reliability, sharpness, spread-error relationships
- "Scale issues" important to consider
  - At what scales should verification be performed?
  - At what scales do the models have skill?
  - At what scales should probabilistic forecasts be presented?

More issues…

- What observational datasets will be used for verification?
  - e.g., Rotation tracks from WDSSII for mesocyclones. MESH (calibrated via Shave) for Hail.

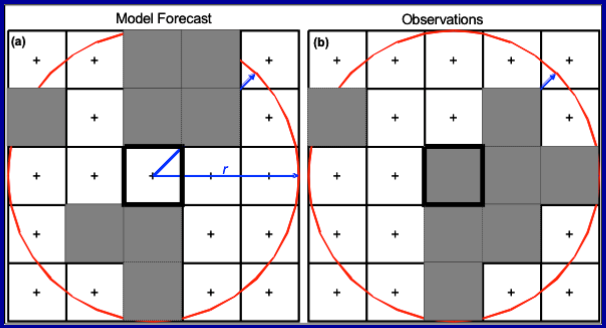
- How to test for statistical significance. Not easy! Need to take into account spatial/temporal autocorrelation of errors.
  - Resampling (e.g., Hamill 1999)
  - Field significance (Elmore et al. 2006)

- What model fields are needed and how frequently should they be output?
  - Use hourly-max fields to save time/space?

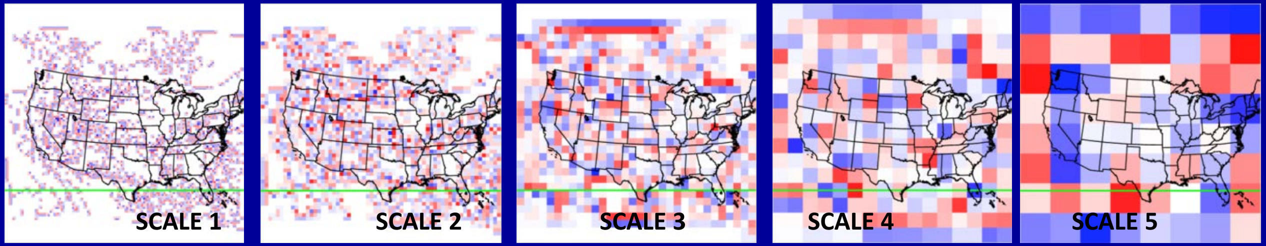
- Efficient methods to quickly visualize distributions of forecast storm attributes are needed.
  - For example, a forecaster should be able to quickly assess number of ensemble members that forecast long track and/or intense mesos. Or, whether forecast PDF is bimodal – some members break the cap and some do not.

Non-traditional methods for verifying WoF

- Neighborhood methods – Consider neighborhood around each grid-point to compute various metrics.



- Scale separation – Examine spatial error field at different scales using wavelets.



Object-based methods for verifying severe storms

- Objects defined as contiguous regions of observed/model grid-points exceeding predefined threshold. Object attributes like location, size, and shape can be compared.
- 2-D object-based algorithms have been around for a while – e.g., MODE (Method for Object-based Diagnostic Evaluation; Davis et al. 2006) – and many useful applications have been illustrated.
- Lack of 3rd dimension limits ability to track time evolution of objects – time evolution of storms is what we are most interested for WoF (e.g., storm tracks, duration, speed, etc.)!
- DTC/NCAR will be releasing "MODE-TD" soon. How easy this will be to use with extremely high resolution forecasts? For preliminary testing, I have worked with MODE-TD and codes that do similar things as MODE-TD for several applications:
  - 1) Measuring track lengths and maximum intensity of simulated rotating storms.
  - 2) Defining the timing of observed and forecast convective initiation.
  - 3) Tracking precipitation systems in CAMs.

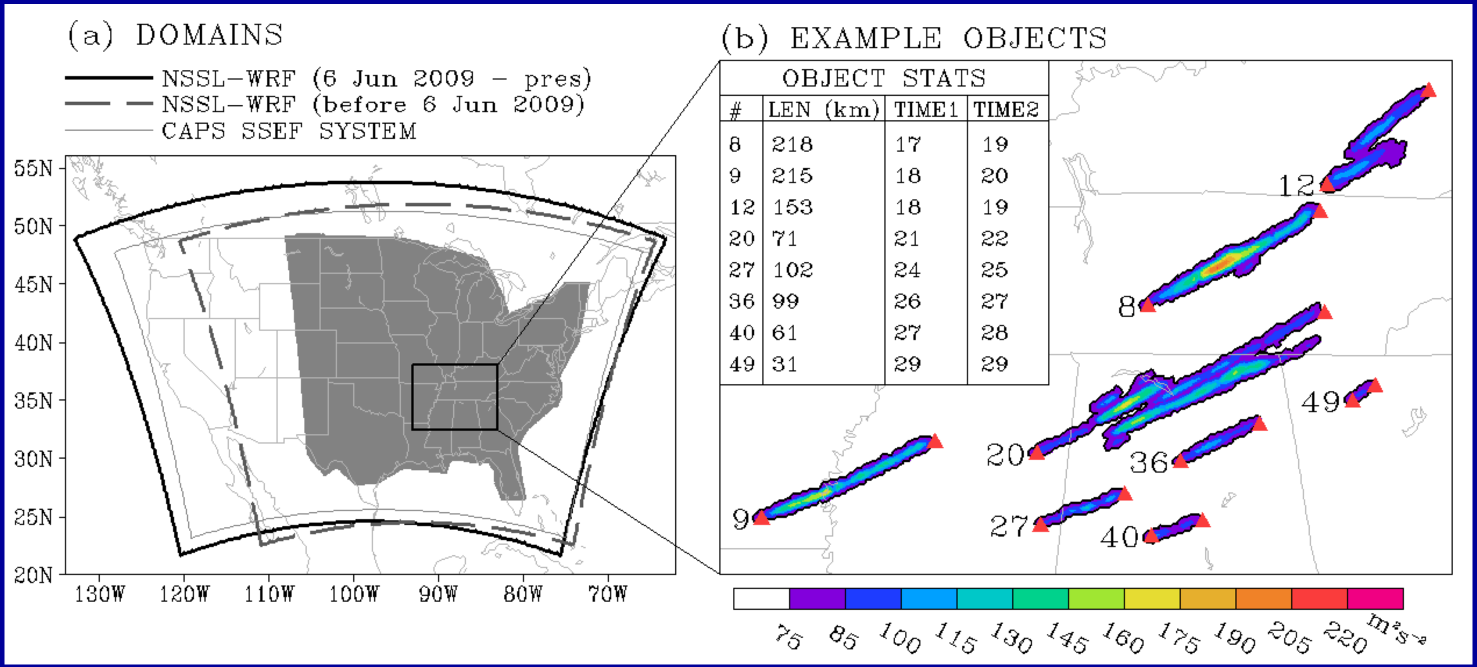
Application 1: Visualization/verification of simulated rotating storm track lengths

- Clark, A. J., J. S. Kain, P. T. Marsh, J. Correia, Jr., M. Xue, and F. Kong, 2012: Forecasting tornado pathlengths using a three-dimensional object identification algorithm applied to convection-allowing forecasts. *Wea. Forecasting*, **27**, 1090-1113.

- Clark, A. J., J. Gao, P. T. Marsh, T. Smith, J. S. Kain, J. Correia, Jr., M. Xue, and F. Kong, 2013: Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Wea. Forecasting*, **28**, 387-407.

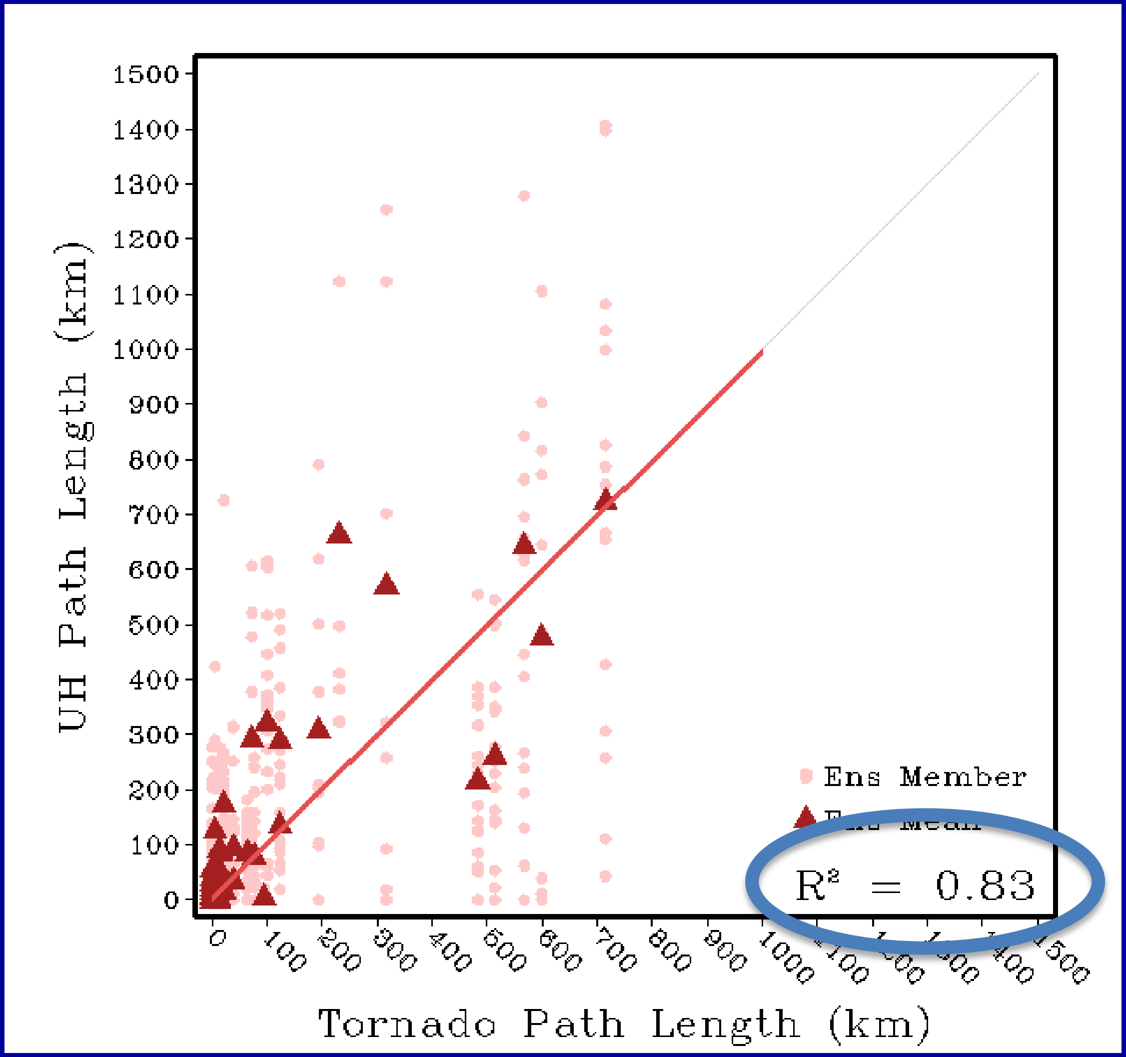
- 3-D object code is applied to hourly-max updraft helicity (UH) to identify number, length, and intensity of 3D UH objects (i.e. rotating storm tracks).

- A study was done on whether total UH path lengths could be used as a proxy for total tornado path lengths (Clark et al. 2012).



Results

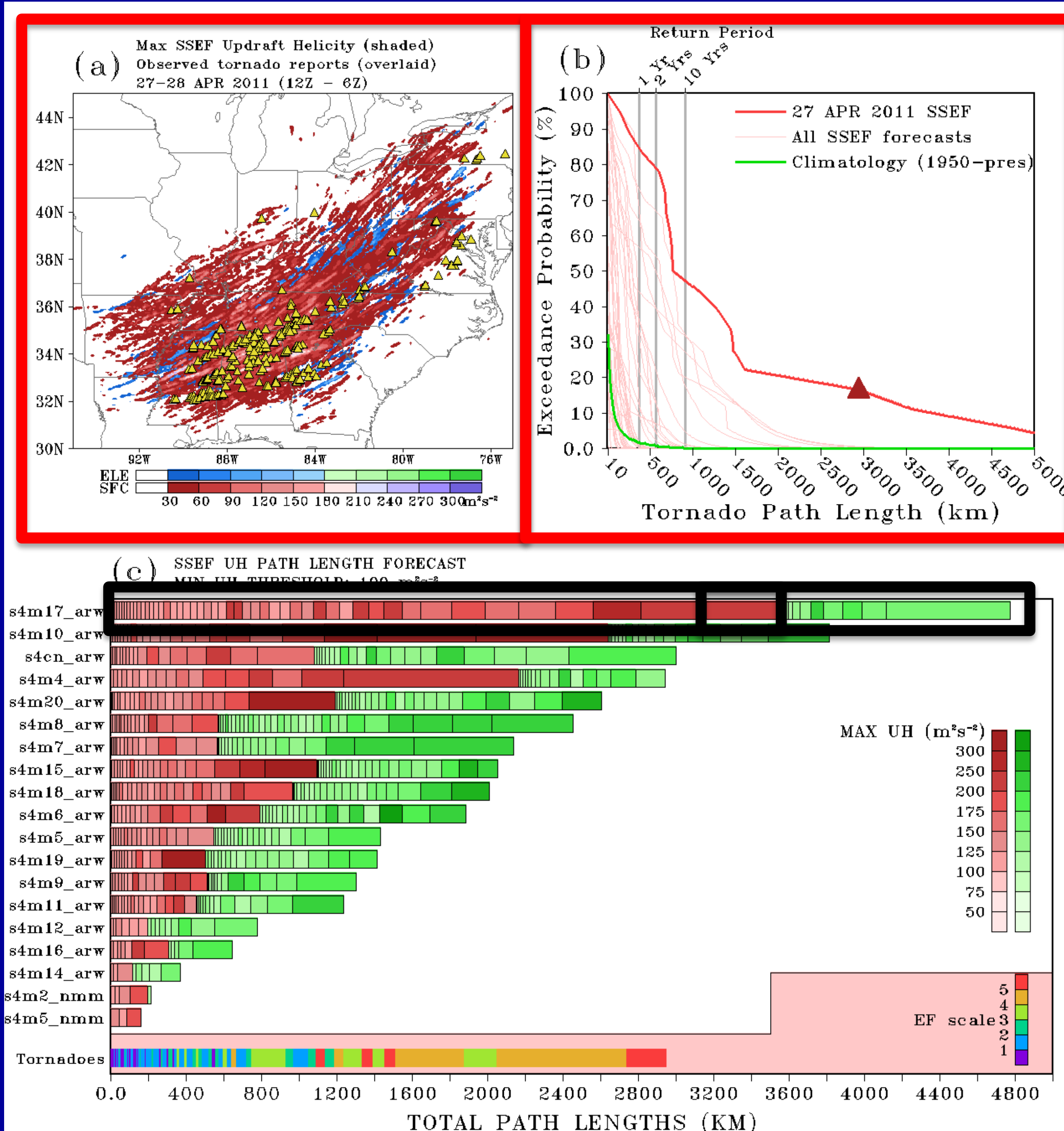
Filtered and calibrated UH > 100 m^2/s^2



- For each case, total track length of 3D UH-objects from each SSEF ensemble member was plotted against the total tornado path lengths for corresponding time periods – UH path lengths identified using a threshold of 100 m^2s^{-2} are shown here because it worked best.

- Portions of tracks from simulated storms that were high-based or elevated, were filtered out. For details… ask later!

- The technique work well, but how do we efficiently present information on 3D UH-objects and utilize inherent uncertainty information provided by the ensemble?



Example UH Forecast Product: 27 April 2011

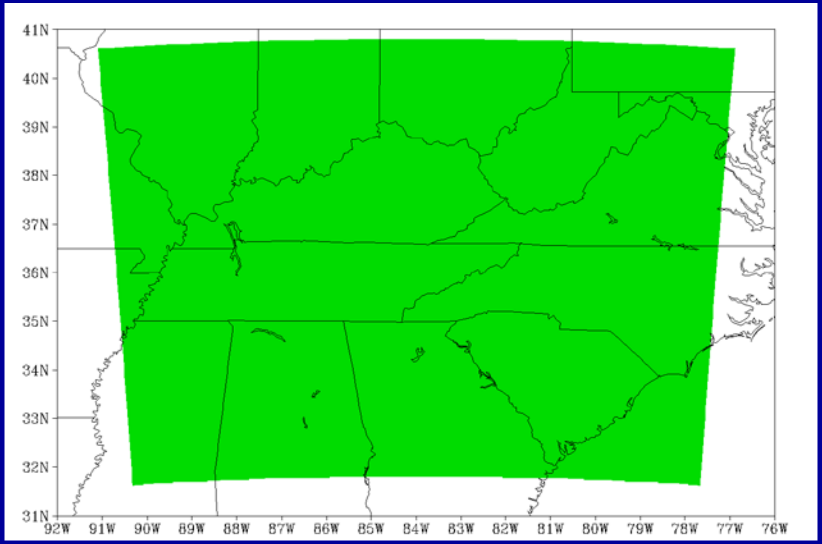
- Length of segments correspond to individual UH tracks. Shading levels denote max intensity of UH within each track. Red shading is for surface based and green for elevated UH tracks.

- Length of entire row is the total UH path length for an ensemble member; members are ordered longest to shortest.

- Thick red line are 27 April exceedance probabilities for total tornado path lengths above 500 km
- Thin red shading for exceedance probs for all other days
- Green line – climo
- Grey vertical lines mark path lengths corresponding to 1, 2, and 10 year return periods

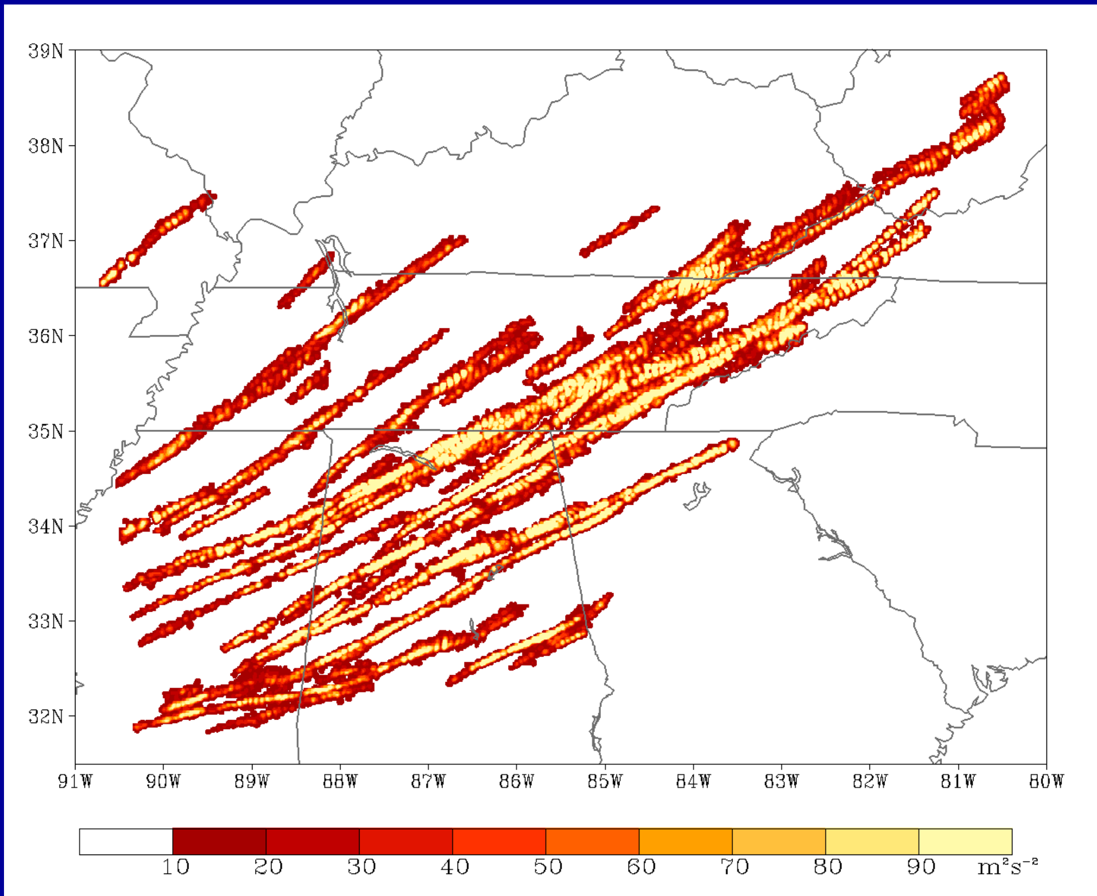
Computing observed UH for 27 April

- A 3DVAR data assimilation system (Gao et al. 2009) used in the 2011 Experimental Warning Program Spring Experiment was run over a domain covering 27 April outbreak.
  - WSR-88D reflectivity and velocity data assimilated at 1.25-km grid-spacing every 5 minutes over the period 15Z to 3Z, 27-28 April – gave 144 separate high-resolution analyses.
  - 12-km NAM forecast valid at analysis time used as first guess background.
  - This 3DVAR system designed for identifying mesocyclones – observed UH easily computed using same formulation as in model.

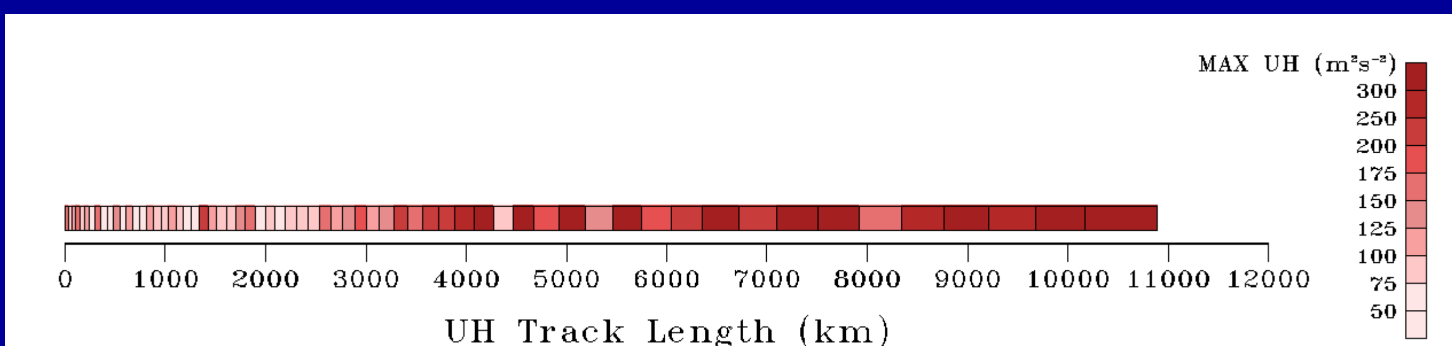


Filtered maximum UH: 1500 to 0300 27-28 April

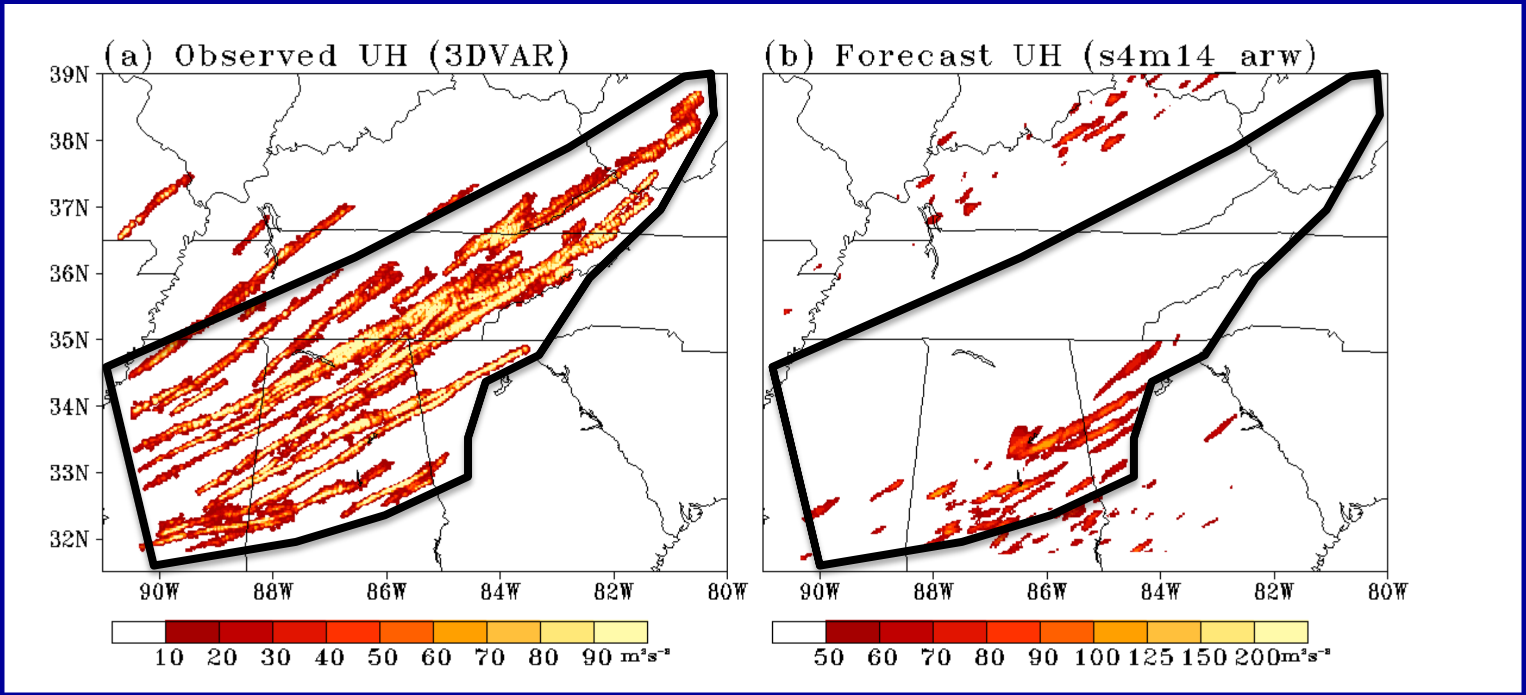
Raw maximum UH: 1500 to 0300 27-28 April

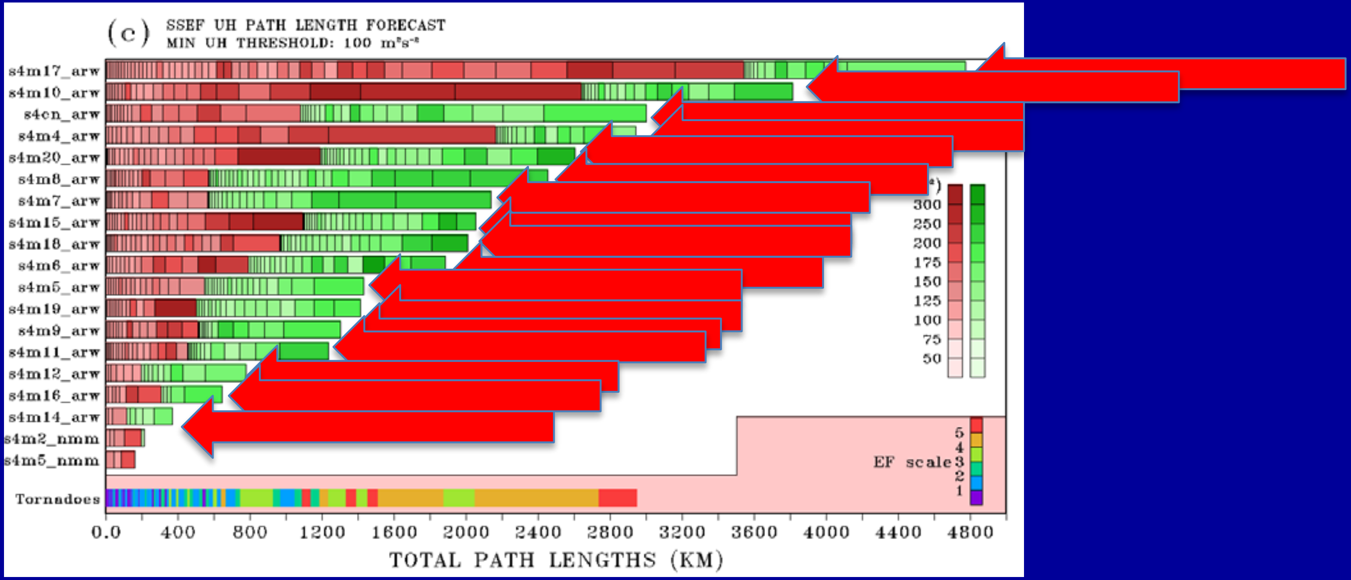


- Apply 3D object algorithm in sensible way and only plot identified objects…
- 64 total UH tracks identified.
- Longest track was ~ 725 km and 12 tracks were over 300 km long.
- Total rotating storm path length of almost 11000 km.



Observed UH compared to individual SSEF members

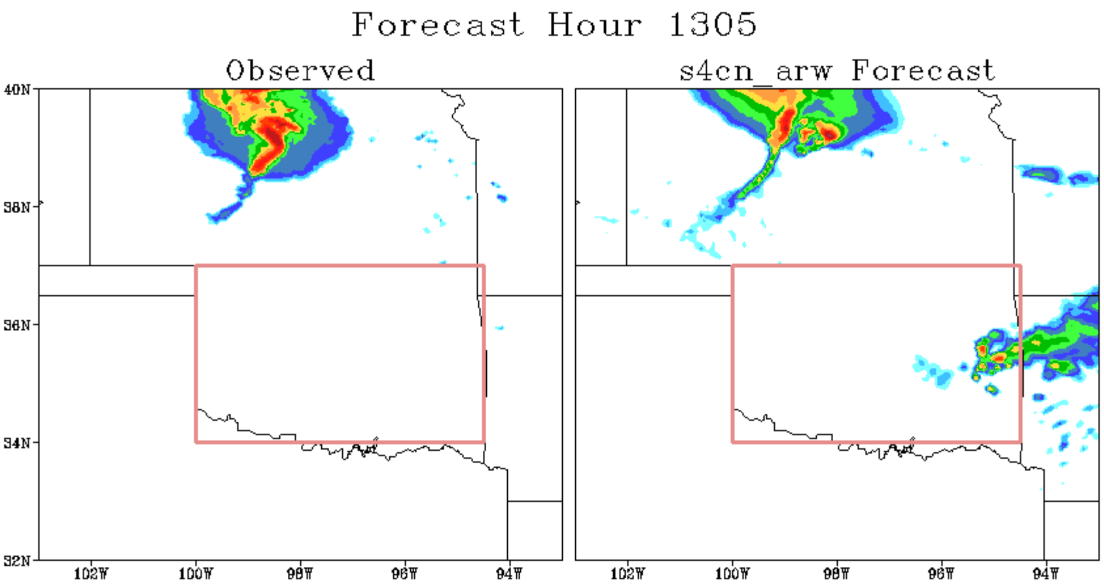


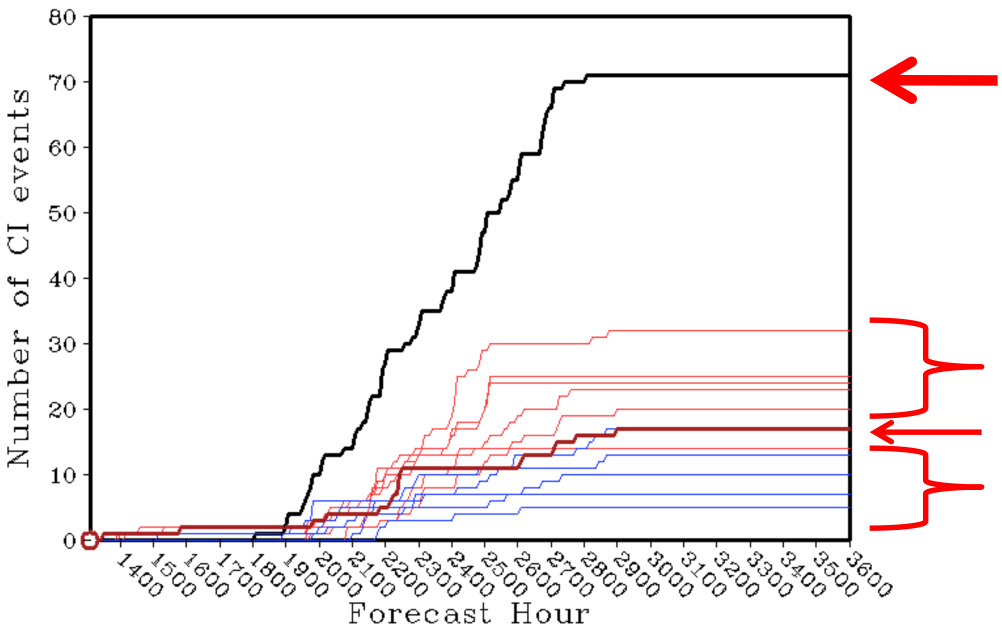


Application 2: Convective Initiation

- Kain, J. S., and Coauthors, 2013: A feasibility study for probabilistic convection initiation forecasts based on explicit numerical guidance. *Bull. Amer. Meteor. Soc.*, **94**, 1213-1225.
- For CI-component of SFE2011, convective activity (CA) was defined as DbZ > 35 at -10 C level.
- To identify CI events, 3D CA objects can be defined and grid-points within the objects with the earliest time are CI.
- Additionally, other grid-points that are local time minima within 3D objects are identified as CI.
  - This allows "merging storms" to have a unique CI event assigned.

Example:
CI over Oklahoma 24 May 2011





Observed CI: Many more storms than in ensemble members

Microphysics members (MYJ PBL)

Control member (Thompson/MYJ)

PBL members (Thompson MP)

Application 3: Tracking Precipitation Systems in Convection-allowing models

- Clark, A. J., R. G. Bullock, T. L. Jensen, M. Xue, and F. Kong, 2014: Application of object-based time-domain diagnostic for tracking precipitation systems in convection-allowing models. *Wea. Forecasting*, **29**, 517-542.

- For 30 h forecasts, MODE-TD used to identify space-time 1-h accumulated precipitation objects in 4 members of the 2010 SSEF system that had identical configurations except for microphysics parameterization, as well as corresponding Stage IV observations.

- Why?

- During SFE2010, we first began to document differences in convective system depiction/behavior with different microphysics. We noticed some big differences!

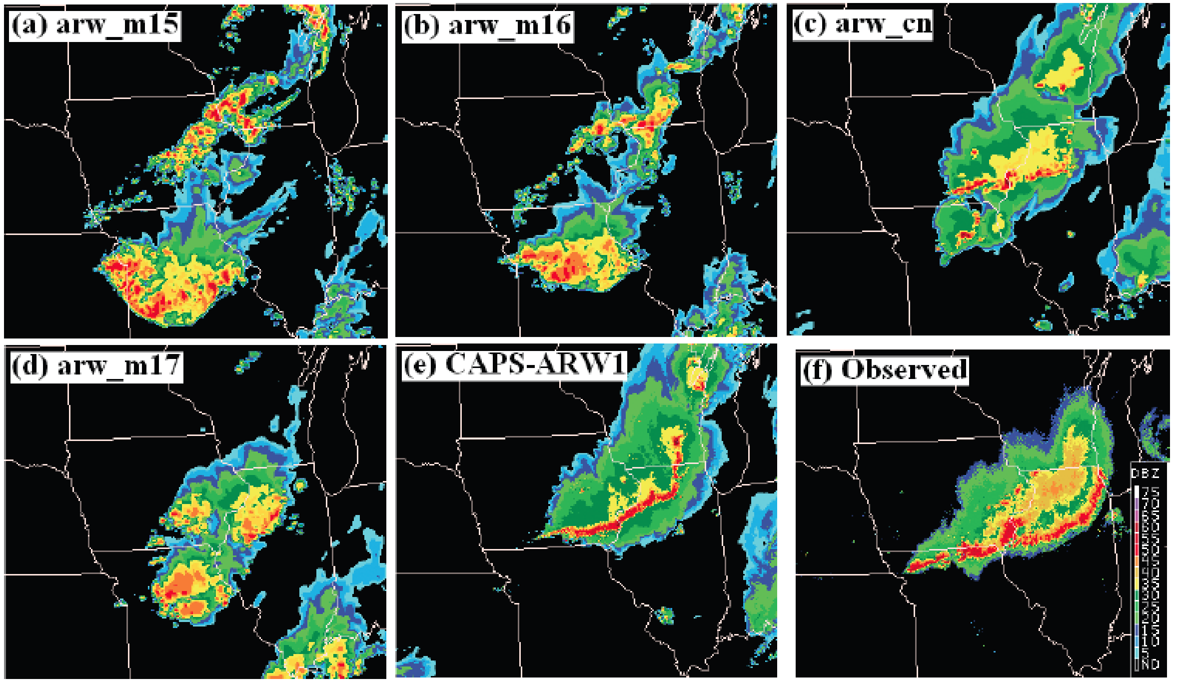
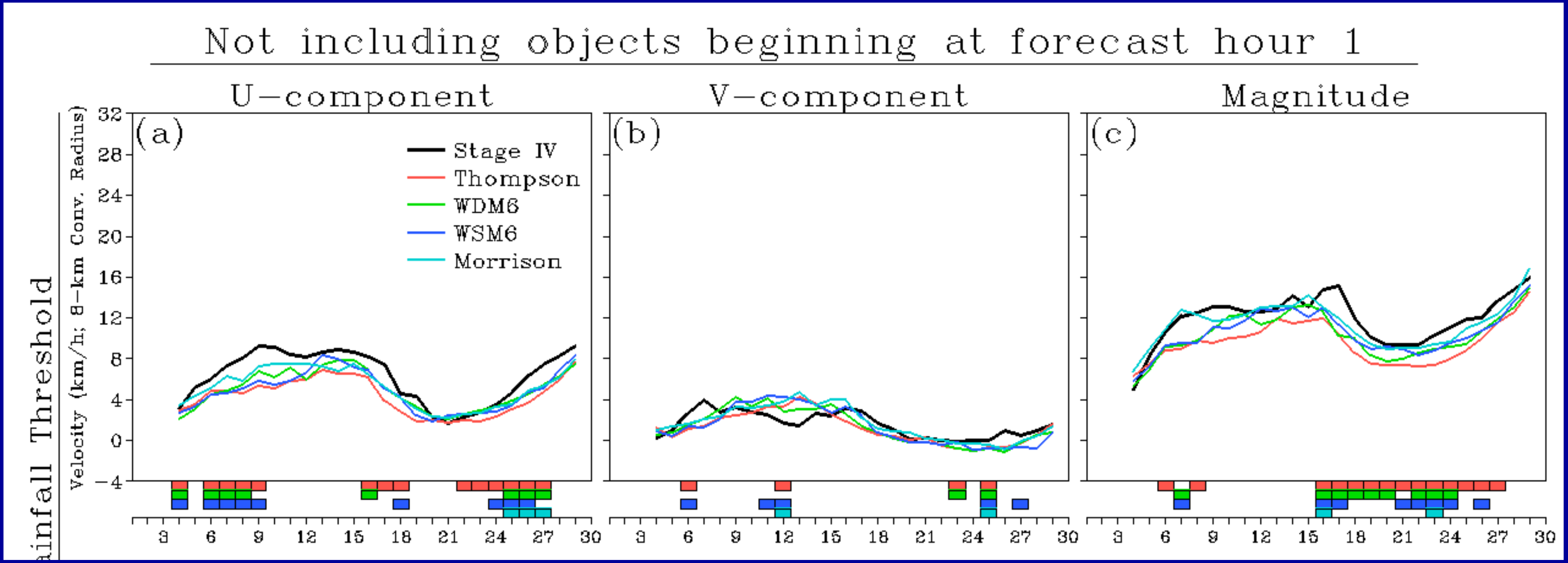


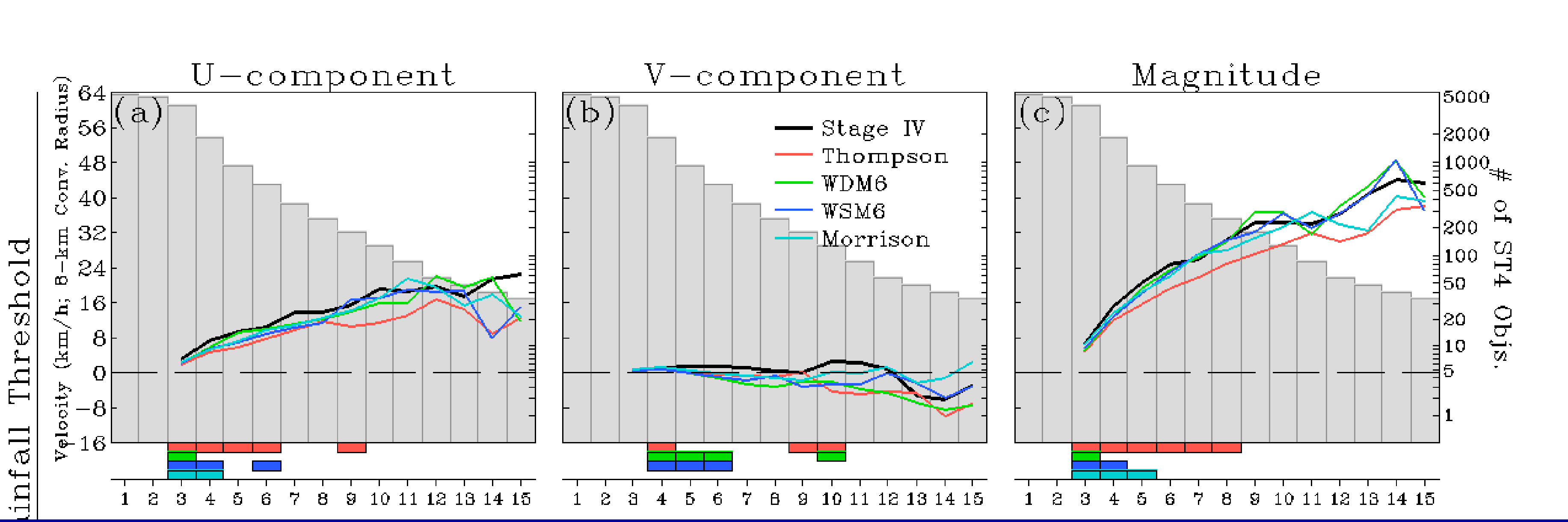
Fig. 3. (a)–(d) Simulated composite reflectivity from SSEF members with identical configurations except for microphysics schemes for 27-h forecasts initialized 0000 UTC 18 Jun 2010. Microphysics schemes are (a) WDM6, (b) WSM6, (c) Thompson, and (d) Morrison. (e) As in (c), but simulated composite reflectivity forecasts are from a CAPS run with 1-km grid spacing. (f) Corresponding observations of composite reflectivity.

- Main results – All schemes too slow, especially during the first part of the forecast.
- Thompson overall is slowest (red line).
- Slow bias at beginning, likely due to inability of the 3DVAR system to properly depict the mesoscale dynamics driving the movement of convective systems existing at the model initialization time.
    - Can be confirmed by comparing speeds excluding objects beginning at forecast hour 1.

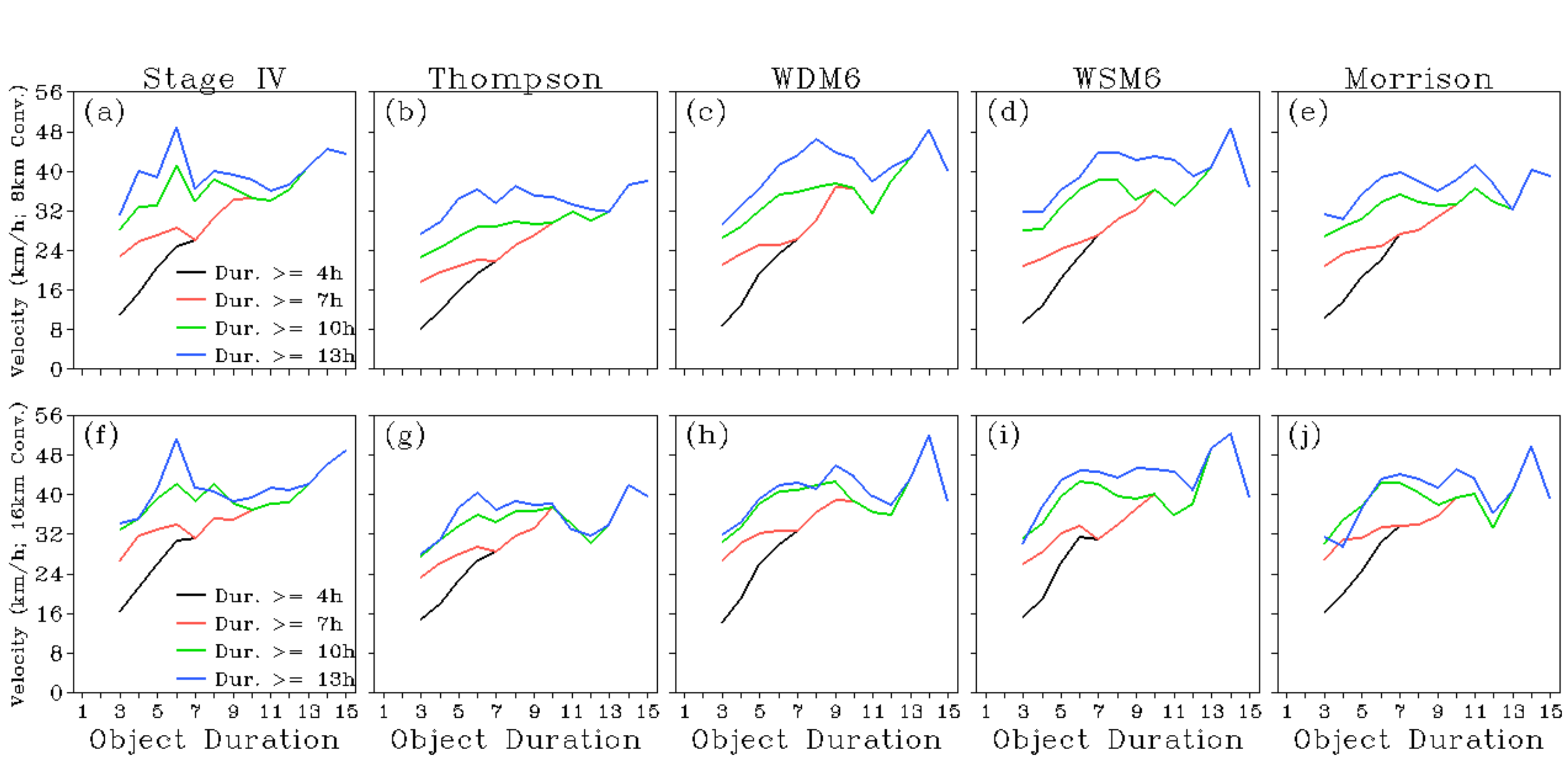


Other results…

- Average velocity components were also computed over the lifetime of time-domain objects – i.e., the start time of each object was set to a common hour and then averages were computed at each subsequent hour.
- Thompson slowest again.
- Objects accelerate with time – perhaps from discrete storms or multi-cell storm clusters starting off moving slow then congealing/growing upscale and accelerating?



- Acceleration an artifact of shorter duration objects having slower speeds and more weight during the first few hours of the average object's lifetime.



Conclusions…

- Recently developed methods for quantifying skill of human severe weather outlooks seem to work quite well.
- Verifying NWP forecasts of severe storms requires new/innovative verification strategies. Methods that consider "time" I think have the most potential to give useful information. MODE-TD!
- Questions?