EVALUATION AND VERIFICATION OF PUBLIC WEATHER SERVICES

Pablo Santos
Meteorologist In Charge
National Weather Service
Miami, FL

WHAT IS THE MAIN DIFFERENCE BETWEEN A GOVERNMENT WEATHER SERVICE FORECAST AND A MEDIA FORECAST???

- ***ACCOUNTABILITY!!!***

HOW ARE FORECASTS USED??

- **HOW MEDIA FORECAST ARE USED:**
  - **What Should I Wear Outside?**
  - **Is It Going To Rain Today?**
- **HOW GOV'T FORECAST ARE USED:**
  - **Who Needs To Be Mobilized?**
  - **What Areas Must Be Evacuated**
  - **What Should I Tell The President/ Governor/Prime Minister, Etc…**

VERIFICATION GOALS:

- **To Gage The Accuracy, Skill And Timeliness Of Warnings, Watches And Forecasts**
- **To Provide A Baseline To Assist In Setting Goals For Measuring Performance**
- **Identify Training Needs To Assist In Improvement Of Forecast And Warning Process & Products**

Evaluation Begins By Establishing Performance Goals

GPRA Fiscal Year Goals

GPRA Metric \ FY Goals	FY 2009 Goal	FY 2010 Goal	FY 2011 Goal	FY 2012 Goal	FY 2013 Goal	FY 2014 Goal	FY 2015 Goal	FY 2016 Goal
Tornado Warnings Lead Time	12	12	12	13	13	13	13	N/A
Tornado Warnings Accuracy	69	70	70	72	72	72	72	N/A
Tornado Warnings False Alarm Ratio	72	72	72	70	70	70	70	N/A
Flash Flood Warnings Lead Time	49	49	49	49	49	49	49	N/A
Flash Flood Warnings Accuracy	90	75	75	76	76	76	76	N/A
Marine Wind Speed Forecast Accuracy	69	69	69	70	70	70	71	N/A
Marine Wave Height Forecast Accuracy	74	74	74	75	75	75	76	N/A
Aviation Forecast IFR Accuracy	64	65	66	67	68	69	69	N/A
Aviation Forecast IFR False Alarm Ratio	43	42	41	40	39	38	38	N/A
Winter Storm Warnings Lead Time	15	15	16	16	16	17	17	N/A
Winter Storm Warnings Accuracy	90	90	90	91	91	91	91	N/A
Precip Forecast Day 1 Threat Score	29	30	30	31	32	32	33	N/A
US Seasonal Temp Forecast Skill	23	24	24	25	25	26	26	N/A
Hurricane Forecasts Track (48 hrs)	108	107	106	102	100	98	96	N/A
Hurricane Forecasts Intensity (48 hrs)	13	13	13	12	12	12	12	N/A

REGARDLESS OF THE USER… ANY FORECAST SHOULD BE EVALUATED IN 3 WAYS:

- **Is It Timely And Accurate?**
- **Is It Understandable?**
- **Does It Meet The Users Needs?**
- **The First Goal Can Be Objectively Evaluated… The Others Take Work!**

OBJECTIVE VERIFICATION

- **Public Forecasts & Warnings**
- **Aviation Forecasts**
- **Marine Forecasts & Warnings**
- **Fire Weather Forecasts**
- **Goal Is To Provide Effective Objective Verification Of All Forecast Parameters**

PUBLIC FORECASTS

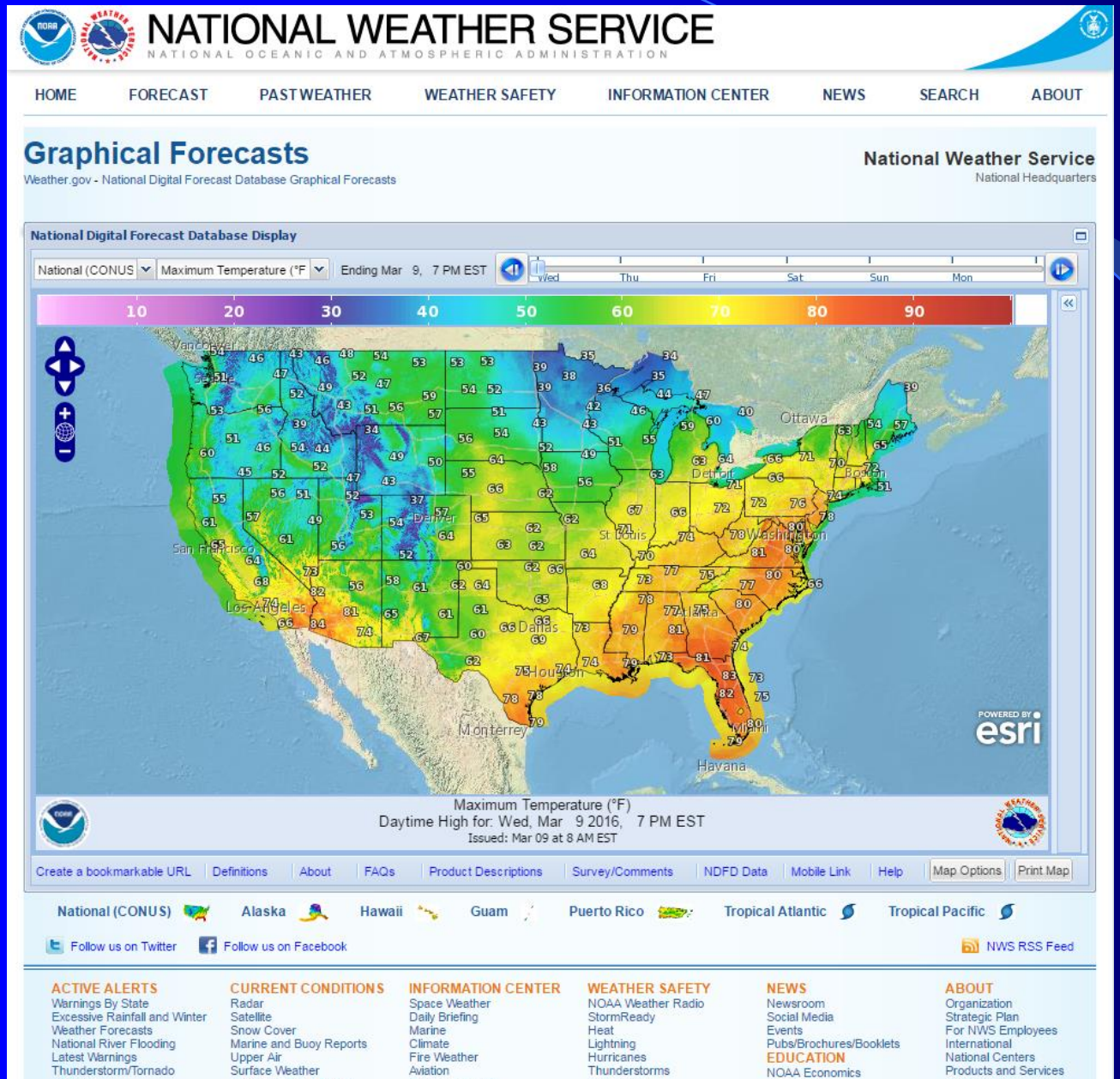
- **Primary Forecast Elements Include:**
  - **Maximum & Minimum Temperatures**
  - **Probability Of Precipitation**
- **Secondary Elements Include:**
  - **Sky Condition (Cloudy, Partly Cloudy, Etc.)**
  - **Winds**
  - **Precipitation Type (Liquid vs. Frozen, Etc.)**

PUBLIC FORECASTS

- **While All NWS Products Are Produced In Text Format…**
- **Those Text Forecasts Are Derived From "Gridded" Forecasts Of Individual Forecast Parameter Fields…**

All Forecast Elements Found in

http://digital.weather.gov/



Gridded, Or Graphical Forecasts Can Be Much More Useful In Depicting Changes On A Spatial Or Temporal Scale…

…But It Makes Realistic Verification Much More Difficult!

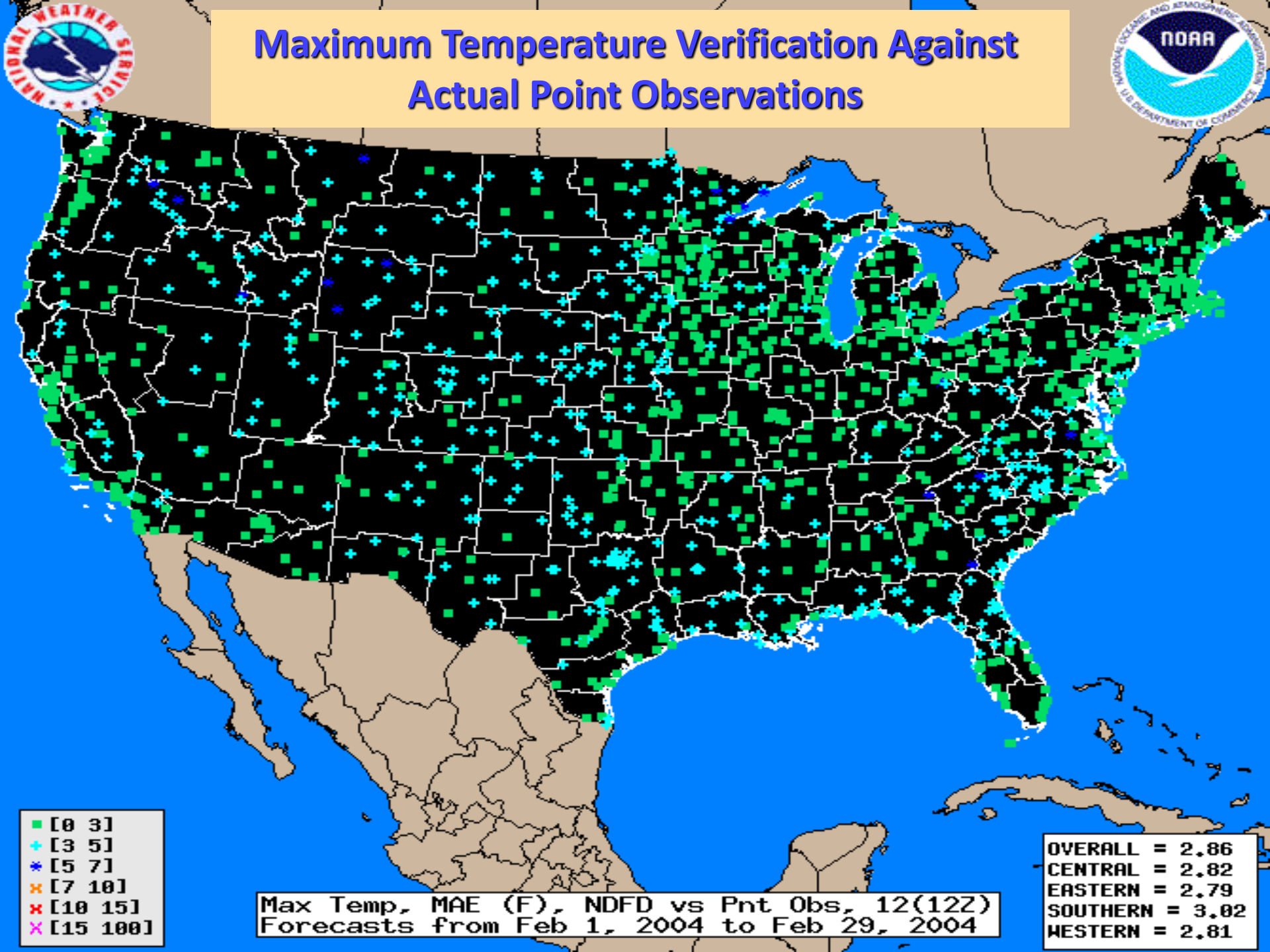
HOW DO WE EFFECTIVELY VERIFY GRIDDED FIELDS?

THE FIRST QUESTION THAT MUST BE ANSWERED IS… WHAT ARE WE GOING TO VERIFY *AGAINST*???

- ***Actual Observations (Points)?***
- ***Model Forecast Fields?***
- ***Data Assimilation Fields?***

Any of these options have their own set of problems!

Maximum Temperature Verification Against Actual Point Observations



Verification Against Actual Point Observations

Strengths:

- Probably The Most Realistic Scheme… Uses Data People Understand
- Deals With Absolute Error… Therefore Is Easiest To Use and Understand…

Limitations:

- Does Not Really Measure "Skill"… No Control To Measure Against
- Tends To Be "Biased" In Favor Of Locations With Smaller Ranges Of Conditions
- Often, The "Point" Location Where The Observation Is Taken Is Not Representative Of The Area Where The Population Lives!

Let's Look At This "Bias" Issue And What It Means In A Comparison Of Temperature Forecast Verification Data For Two Sites…

Site A:

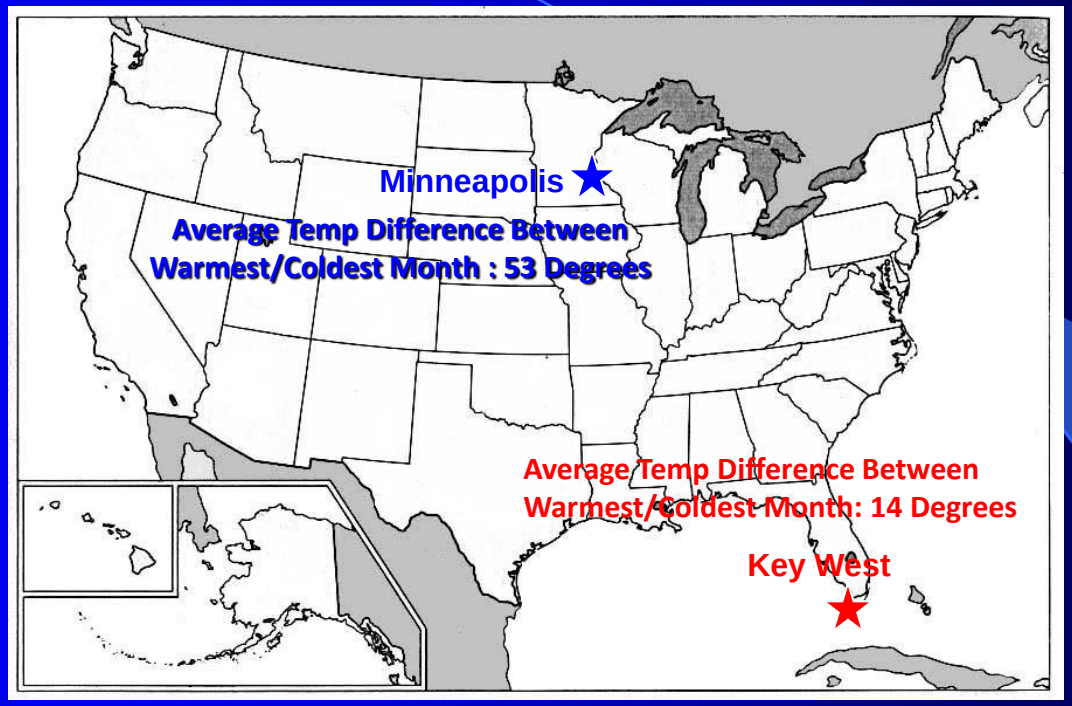
- *94% Of Temp Forecasts Have Errors Of Less Than 5 Degrees…*
- *Less Than 1% Of Forecasts Have Errors Of 10 Degrees Or More*

Site B:

- *70% Of Temp Forecasts Have Errors Of Less Than 5 Degrees…*
- *8% Of Forecasts Have Errors Of 10 Degrees Or More*

Are The Site A Forecasters Really That Much Better Than Those At Site B? *Should We Fire The Forecasters At Site B???*

First, Let's Make Sure That This Is Actually A Fair Comparison!
Site A Is Actually Key West… Site B Is Minneapolis



The Answer To Our Question: Nope… Key West Simply Does Not Have As Active A Weather and Climate Regime as Minneapolis… Therefore, Fewer and Smaller Weather Changes

This Means That Using This Type Of Absolute Comparison For All Forecast Offices Does Not Provide A Realistic Evaluation Of Forecast Skill

Is It Representative? Consider Jacksonville, FL

Verification Against Model, or Numerical Forecast Fields

Strengths:

- Easier To Judge Actual "Skill" Since Both Model and Human Forecasts Are Compared Against A Common Parameter… Actual Observations

- Is Much More Effective At Providing A Realistic Evaluation Of Forecaster Performance…

Limitations:

- In The Public's Eyes… The "Comparison" Isn't Much Of A Factor

- If A Temperature Forecast Is Off by 20 Degrees, It's Not Much Comfort To The Public To Know That The Model Missed It By 25 Degrees!

Let's Look At This Verification Scheme For The Two Sites We Used Previously …

Key West:

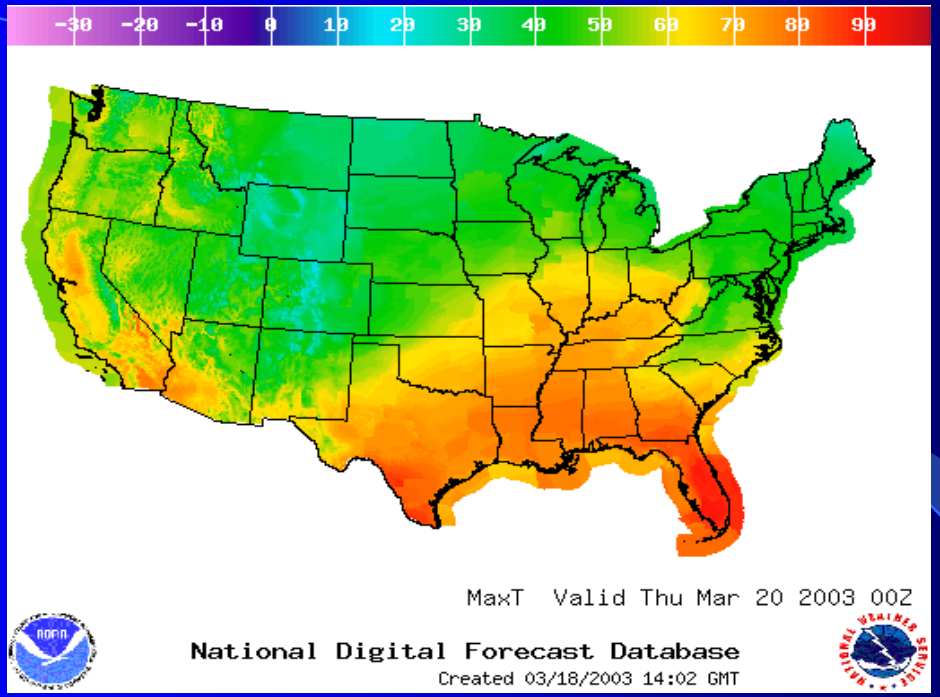
- Recall That 94% Of Temp Forecasts For This Station Have Errors Of Less Than 5 Degrees… But The Raw Model Forecasts Were Actually 0.1% Better!

- In Addition… The Model Guidance Had 10% Fewer Errors of More Than 5 Degrees

Minneapolis:

- While The Absolute Errors For This Location Were Greater Than Those For Key West… They Actually Improved Overall Against The Model Forecasts by 2.3%

- And… They Had 4% Fewer Errors of More Than 5 Degrees!



No Matter What Method Is Chosen… Effectively Verifying Forecasts Involving *Dozens Of Fields*… Out To As Much As *7 Days*… At Time Scales As Small As *One Hour*… While Ensuring *Consistency* With Surrounding Offices… Is A Huge Challenge !!

PUBLIC FORECASTS

- **Primary Forecast Elements Include:**
  - **Maximum & Minimum Temperatures**

MAX & MIN TEMPERATURE FORCASTS

- **Verification Efforts Focus On The Actual Value**

- **They Don't, However, Take Into Account:**
  - **What Time The Max Or Min Occurred**
  - **How The Temperature Changed During The Course Of A Forecast Period.**

MAX & MIN TEMPERATURE FORECASTS

- **Consider A Forecast For The Next 12 Hours, Released At 6 Am Which States Or Depicts The Following:**

- **Today: Cloudy. High Of 70.**

- **Actual Conditions: High During The 12 Hours Was 70 And Cloudy Conditions Existed For Most Of The Day.**

- **Did This Forecast Verify?**

MAX & MIN TEMPERATURE FORECASTS

- **From A "Verification" Standpoint This Was An *Accurate* Forecast**
- **However… While The Temp At 6 AM Was 70… A Cold Front Passed Through And The Temp Dropped To Below 50 By 8 AM And Then Spent The Rest Of The Day In The 40s.**
- **It's Highly Debatable, Therefore, That This Would Be Considered A "*Good*" Forecast By Users!**

PUBLIC FORECASTS

- **Primary Forecast Elements Include:**
  - **Maximum & Minimum Temperatures**
  - **Probability Of Precipitation**

PROBABILITY OF PRECIPITATION FORECASTS

- **Probably The Least Understood Forecast Parameter Used By The U.S. NWS.**
- **Originally Designed As A Product Of 2 Probabilities:**
  - **"Conditional" Probability**
  - **"Areal" Probability**
- **Ultimately meant to convey frequency of occurrence given present conditions.**

PROBABILITY OF PRECIPITATION FORECASTS

- **Today's Forecasts Are Verified Against Probabilities Derived From Model Output Statistics… A More Objective… But Very Different Approach.**
- **The Point Could Legitimately Be Made That We Are Comparing Apples And Oranges!**
- **Despite This…The Main Verification Tool Used To Evaluate PoP Forecasts is Brier Score**

Brier Score

- **Measures Accuracy Of A Set Of Probability Assessments (but it says nothing about reliability):**

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

N = Number Of Forecasts,
Ft = Probability That Was Forecast
Ot = Actual Outcome (0 If No Rain, 1 If Rain Occurs)

Brier Score

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

With Brier Score, The Lower Score = Higher Accuracy:

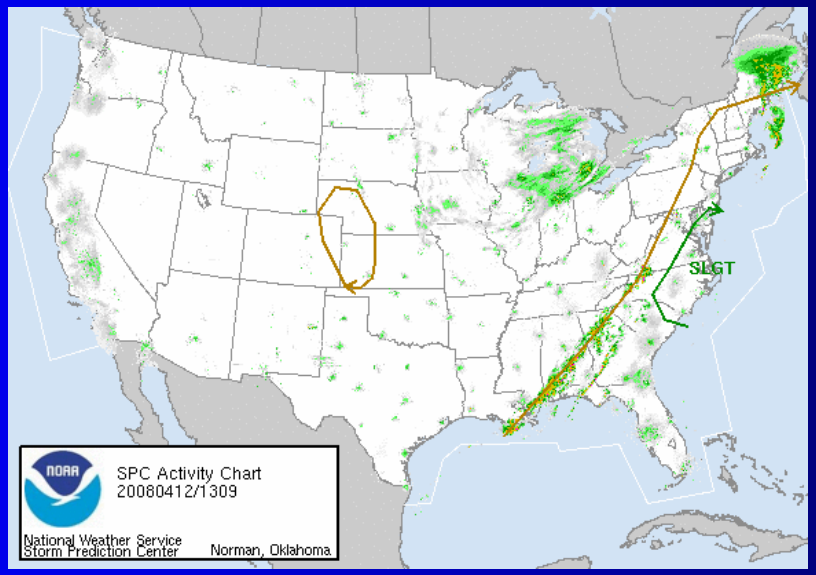
- If you forecast 100% and rain occurs, (ft and ot = 1) BS=0= Perfect!
- If you forecast 100% and no rain, (ft=1, ot=0), BS=1 Awful!
- If you forecast 70% and rain occurs (ft=.7, ot=1) BS=.09 Pretty Good!
- If you forecast 30% and rain occurs (ft=.3, ot=1) BS=.49 Not Too Good!

Brier Score Rewards The Aggressive and Punishes “Hedge” Forecasts!
- If you forecast 50%, your BS=.25 Whether or not rain occurs.
- Since The NWS Performance Goal for PoP is ,<.10, This is Bad.

NOTE: Interestingly enough hedge forecasts are punished by Brier Score yet they might be reliable.

So… Why Not Just Forecast Either High Or Low Probabilities?

Public Perception Is The Big Problem! Consider This Forecast Scenario:



A Fast Moving Cold Front Is Forecast To Move Through…With A Solid…But Narrow Band Of Precipitation. The Forecaster Is Sure It Will Rain So She Forecasts a PoP of 100%

Was It A Good Forecast?

When The Front Passed Through, It Rained For 10 Minutes From 10:00 AM to 10:10 AM Dropping A Total Of .02 Inch.

For Verification Purposes, This Would Be A "Perfect" Brier Score Of 0…. But How *Useful* Was It?

- How About The Afternoon Picnic That Was Cancelled… Even Though The Afternoon Was Totally Dry!

- Or Work Lost On Construction Which Cancelled Work For The Entire Day… Even Though It Only Rained For 10 Minutes!

While Brier Score Is An Effective Statistical Tool For Evaluating Precipitation Forecasts…

12-hr Periods	LCL	1	GUI	LCL	2	GUI
# Forecast Periods	2,917,343			2,925,270		
# Precipitation Cases	542,307			543,750		
Obsvd Precip Frequency	0.186			0.186		
Mean PoP Forecast	0.23		0.23	0.22		0.22
Mean PoP Forecast w/ Precip	0.61		0.59	0.57		0.56
Mean PoP Forecast w/o Precip	0.14		0.14	0.14		0.15
Brier Score (BS)	0.0870		0.0851	0.0915		0.0891
% Imp GUI	-2.3			-2.7		
Brier Score w/ LCL PoP >= 30%	0.2039		0.1926	0.2076		0.1964
% Imp GUI w/ LCL Pop >= 30% (# Cases)	-5.9		(956,119)	-5.7		(943,754)
Brier Score w/ Precip	0.2220		0.2428	0.2586		0.2680
% Imp GUI w/ Precip (# Cases)	8.6		(542,307)	3.5		(543,750)
Brier Score when Changes to GUI >= 20%	0.2221		0.2045	0.2242		0.2025
% Imp GUI w/ Change >= 20%	-8.6		(348,031)	-10.8		(329,209)
Percent of Forecasts Correct	87.7		88.1	87.0		87.5

It Really Doesn't Measure How Helpful The PoP Forecasts Are To The People That Use Them or how reliable they are!!

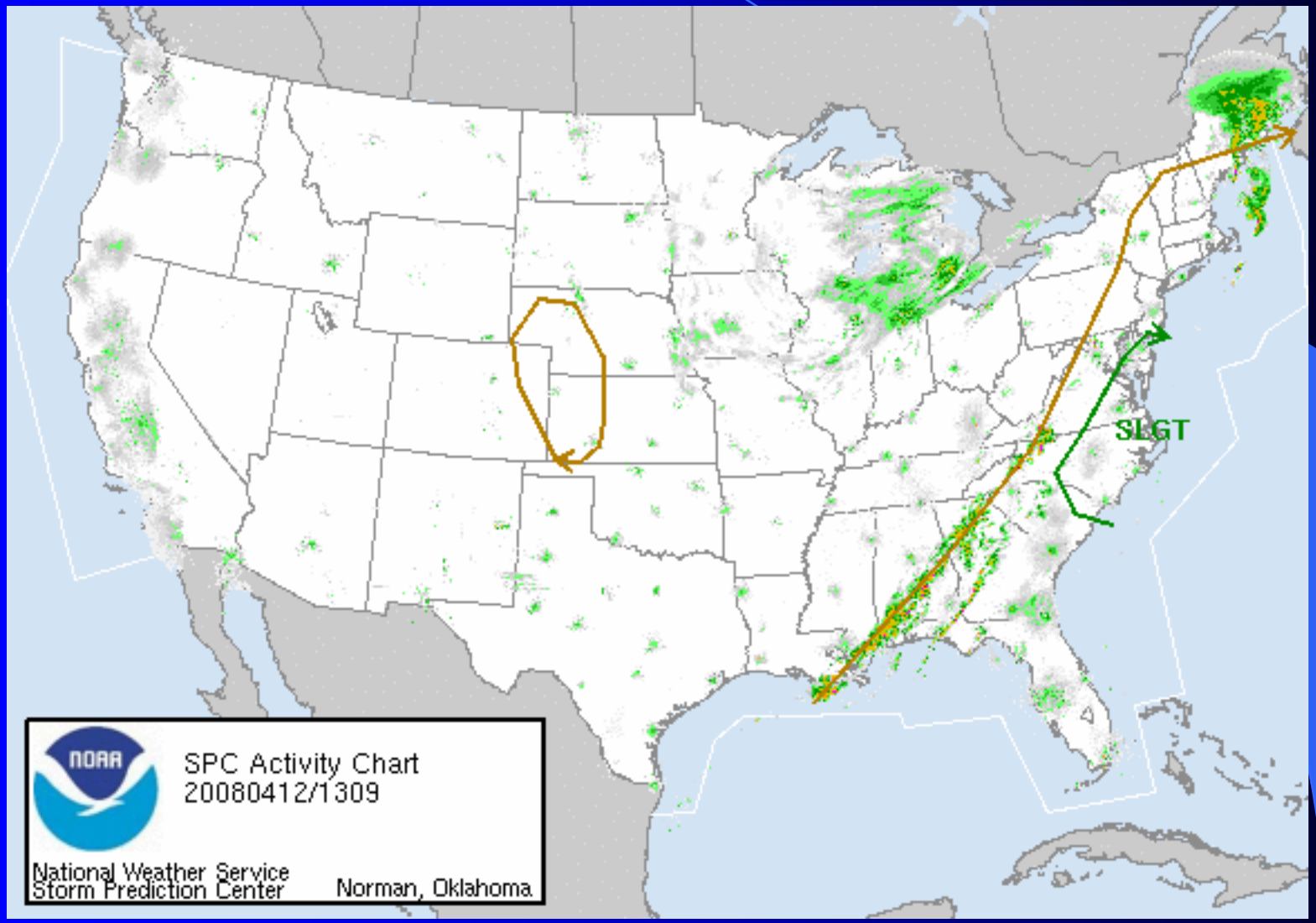
SEVERE LOCAL STORM PRODUCTS AND VERIFICATION

IDEALLY, THESE PRODUCTS FOLLOW A LOGICAL TIME/SPATIAL PATTERN… THREAT AREA AND TIME FRAMES BECOME SMALLER AS THE EVENT BECOMES MORE CERTAIN.

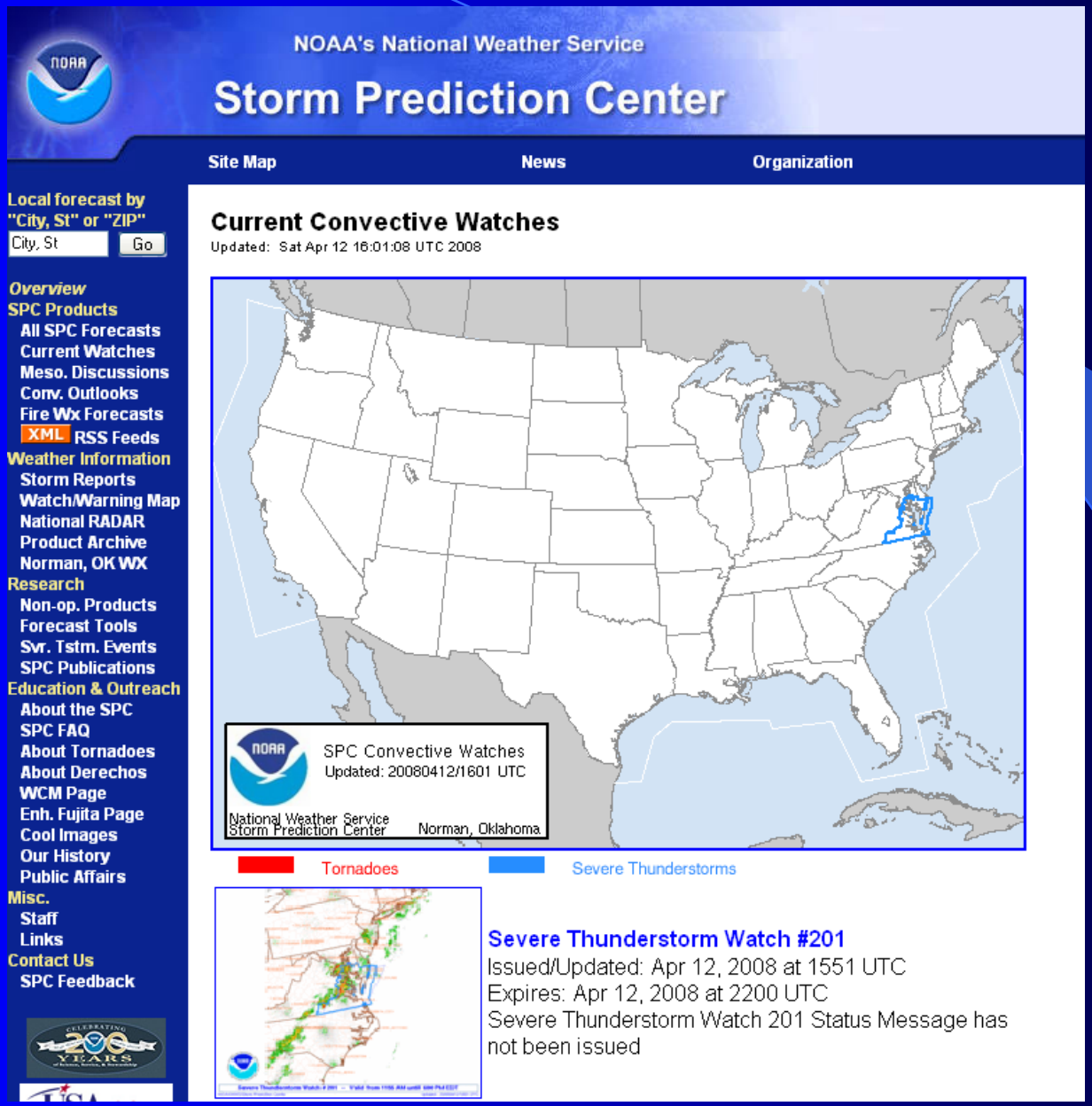
Storm Prediction Center

…Identifies Potential Severe Weather Threat Areas…

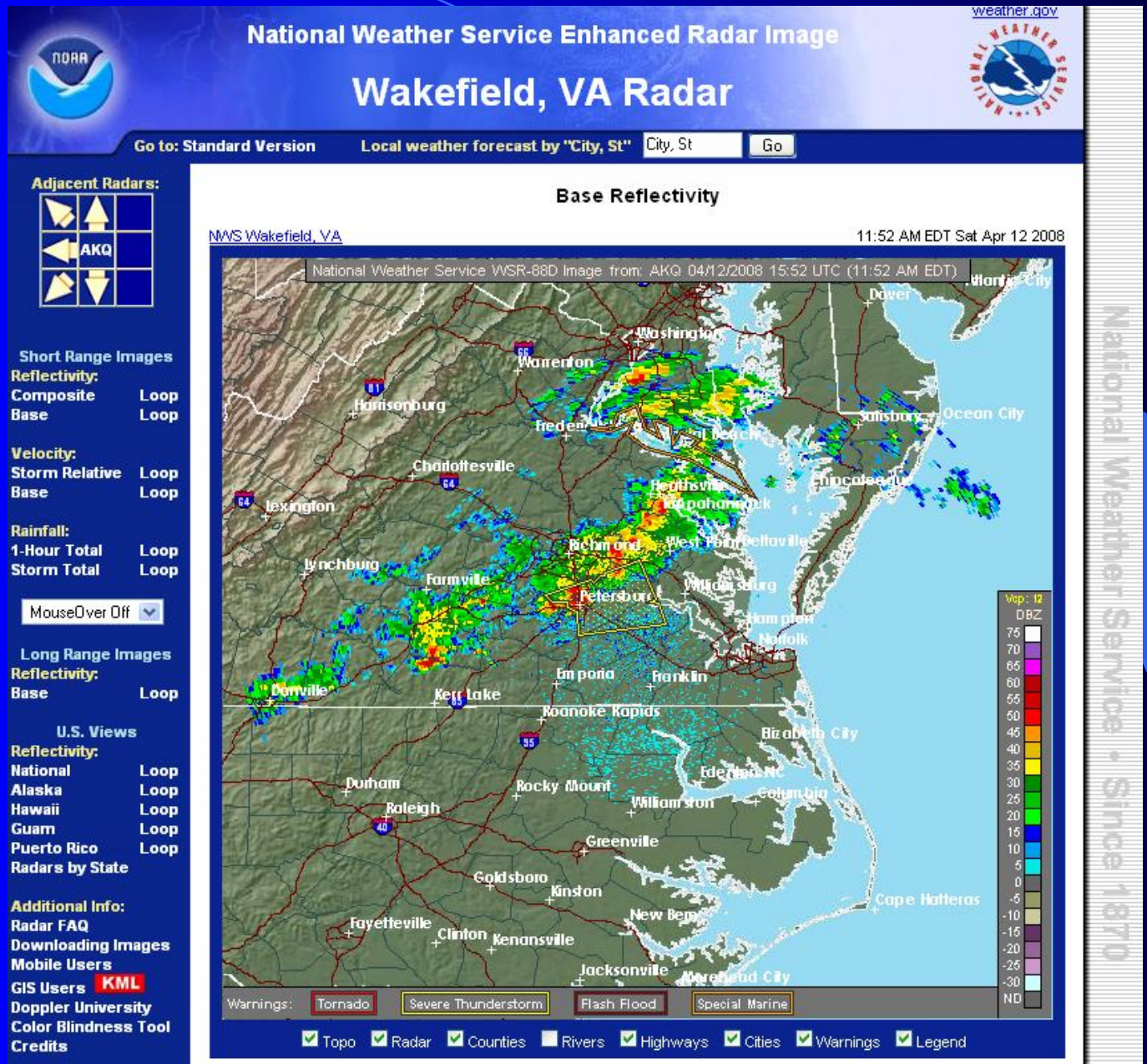
Time Frame: Out As Far As 8 Days



Storm Prediction Center and Local NWS Office Issue "Watches"… Time Frame: Usually 8 Hours Or Less



Local NWS Offices Issue “Storm Based” Warnings : Time Frame: Usually 1 Hour Or Less

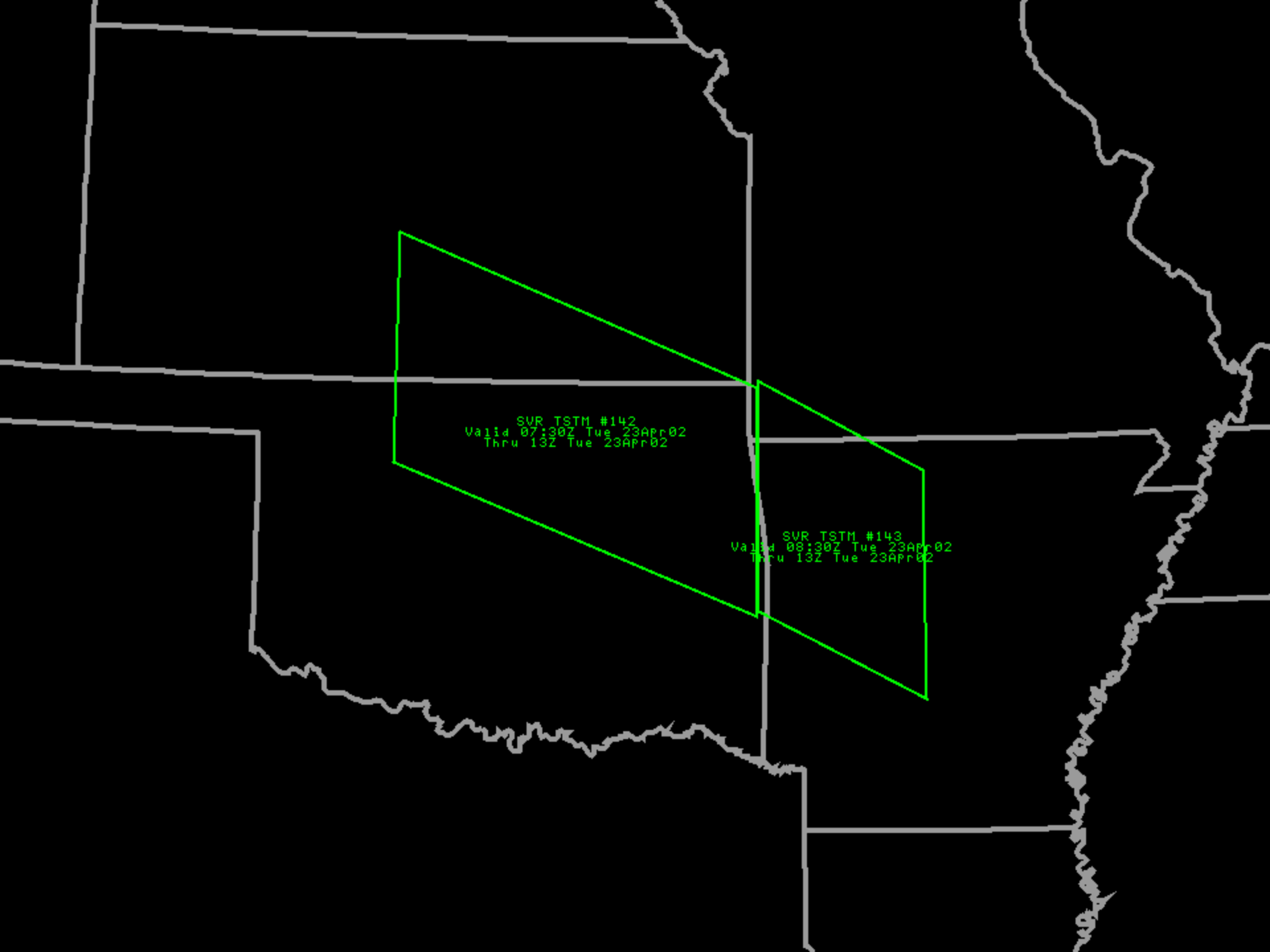


SEVERE WEATHER: WATCHES VS. WARNINGS

- **Watches Indicate Conditions Are Favorable For Severe Weather Development**
- **Warnings Are Issued When Severe Conditions Are Imminent Or Occurring**
- **There Are Important Differences In How These Products Are Verified!**

SEVERE WEATHER: WATCH VERIFICATION

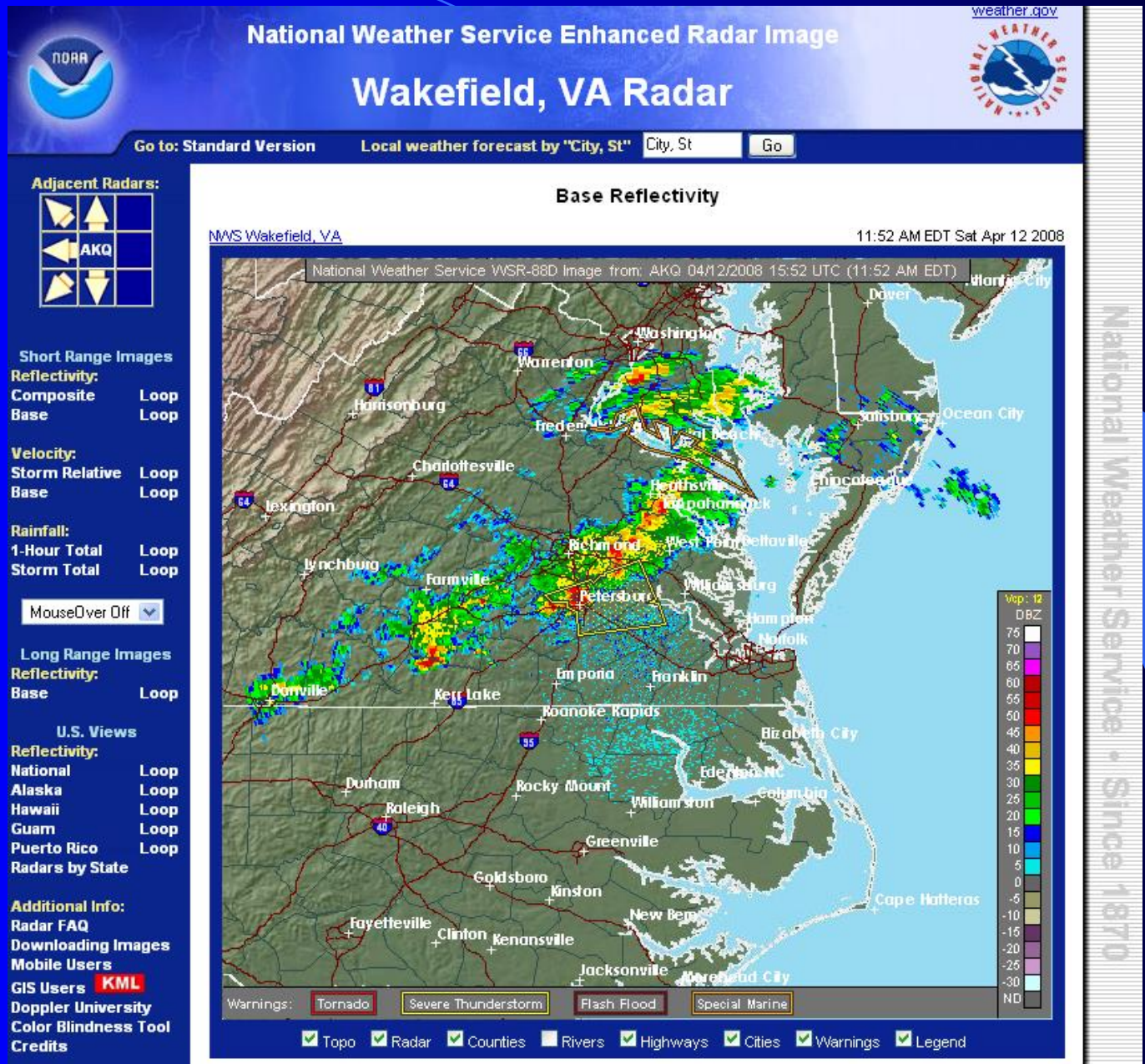
- **Watches Are Generally Issued For Large Geographical Areas… Sometimes Thousands Of Square Miles**
- **A Severe Weather Event Anywhere In This Geographical Area Verifies The Entire Watch**


SVR TSTM #142
Valid 07:30Z Tue 23Apr02
Thru 13Z Tue 23Apr02
SVR TSTM #143
Valid 08:30Z Tue 23Apr02
Thru 13Z Tue 23Apr02

SEVERE WEATHER VERIFICATION: COUNTY VS. STORM-BASED WARNINGS

- Warnings For Severe Thunderstorms Or Tornadoes Used To Be Issued On A County By County Basis.
- However… Storms Do Not Respect Political Boundaries! A Single Storm Will Frequently Threaten Parts Of Several Counties.
- Warnings Are Now Issued As “Polygons” …Based On Specific Storm Location & Motion Rather Than Political Entities.
- This Does Present Some Verification Challenges… But Provides Much More Realistic Verification!

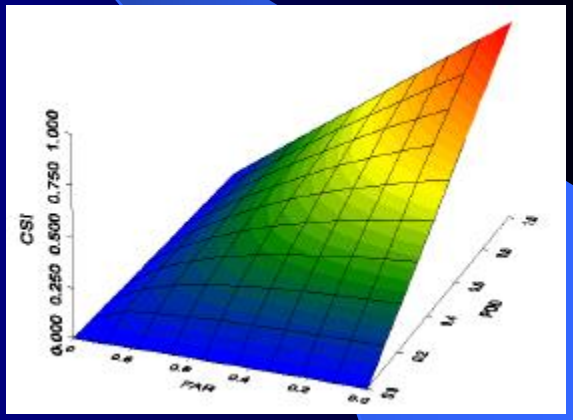
“Storm Based” Warning Example:



SEVERE WEATHER: WARNING VERIFICATION

- Warnings Are Verified Based On Four Primary Computed Parameters:
  - Probability Of Detection (POD = A/(A+C)):
    Was The Event Warned For?
  - False Alarm Ratio (FAR = B/(A+B)):
    Warning Was Issued, Did it Occur?
  - Critical Success Index (CSI = A/(A+B+C))
  - Lead Time Of The Event

2X2 Contingency Table		Event Observed	
		Yes	No
Event Forecast	Yes	A	B
	No	C	D



- Individual Offices Are Responsible For Gathering And Reporting Severe Weather Events In Their Area.
  - This Data Is Published In “Storm Data”

SEVERE WEATHER: WARNING VERIFICATION

- **For Verification Purposes, Severe Weather Includes:**
  - **Thunderstorm Winds Of 50 Knots (58 Mph) Or Greater**
  - **Hail Of 1.00” Diameter Or Greater**
  - **Tornadoes (But Not Funnel Clouds)**
  - **Occurrence Of Structural Wind Damage Which Implies The Existence Of Any Of The Above.**

SEVERE WEATHER: WARNING VERIFICATION

- To Verify A Warning… Event Must Occur Within The Valid Time And Area Of A Warning.
- This Is Subject To Some Complicated Limitations, However.
  - For Example: The "10/15 Rule" – Severe Events Occurring Within 10 Miles Or 15 Minutes Of Each Other Are Considered Duplicates…unless Winds Are 65 Mph Or More Or Hail Is 2" Or Greater… Or It Is The Only Event Verifying A Warning.
  - This Can Obviously Get Quite Confusing.

Summary Statistics

Group	Counts							Statistics							
	Warnings			Events				Scores			Lead Time (min)		Warning Area (sq. mi)		
	Total	Verif	NOT Verif	Total	Fully Warned	Partially Warned	NOT Warned	POD	FAR	CSI	Mean	Initial	Total	Average	County Reduction
National	22747	11766	10981	24219	19633	541	4045	0.824	0.483	0.466	18.91	18.54	15489174.63	681.02	0.83

Summary Statistics

Group	Counts							Statistics							
	Warnings			Events				Scores			Lead Time (min)		Warning Area (sq. mi)		
	Total	Verif	NOT Verif	Total	Fully Warned	Partially Warned	NOT Warned	POD	FAR	CSI	Mean	Initial	Total	Average	County Reduction
JAX	182	127	55	187	174	0	13	0.930	0.302	0.663	16.28	16.24	67988.53	373.56	0.87

WARNING VERIFICATION: "LONG FUSED" EVENTS

- Longer Duration Events Such As Winter Storm Or High Wind Events Are Verified Differently.
  - Main Criteria Is When And If An "Event" As Defined By The NWS Is First Observed In A Warned (Or Unwarned) Area
  - Under Some Circumstances, (Both Long And Short Fused Events) A Warning Issued After The Beginning Of An Event Can Still Be A Verified Warning… But With Zero Lead Time.

SUMMARY

- **While No Verification Scheme Is Perfect, A Good One Will Do The Following:**
  - **Realistically Measure Objective Data In The Way It Was Intended To Be Used**
  - **Compares Forecast Data Sets To Others That Are Truly Forecasting The Same Thing**
  - **Attempts to measure and/or establish usability.**
  - **Can Be Used For Real Time Quality Control To Ensure Forecasts Are Useful To People, And Not Just A Set Of Numbers**
  - **Can Truly Be Used To Identify Trends And Biases And Improve The Forecasts and Warnings**

IMPROVING THE QUALITY OF PUBLIC WEATHER SERVICES

Pablo Santos

Meteorologist in Charge

National Weather Service

Miami, FL

Post-Disaster Service Assessments:

A Key Part Of The Quality Improvement Process!

THREE IMMEDIATE POST-EVENT QUESTIONS:

- **What Went Right?**

- **What Went Wrong?**

- **How Can We Improve?**

Answering These 3 Questions Is The Primary Goal Of A Service Assessment!

WHAT A SERVICE ASSESSMENT IS :

- **A Learning Tool For Future Events**
- **A Way To Identify “Best Practices”**
- **A Way To Identify And Correct Problems**

WHAT A SERVICE ASSESSMENT IS NOT:

- **A Meteorological Study Of The Event**
- **A Historical Documentation Of The Event**
- **A Way To Place The Blame When Things Go Wrong**

THE GOAL OF THE SERVICE ASSESSMENT IS TO:

- **Explain What Happened**
- **Detail NWS Actions Before, During And After The Event**
- **Recommend Changes In NWS Policy, Procedures, Products And Services To Improve Their Quality!**

CRITERIA FOR SERVICE ASSESSMENTS

- **Significant Impact On Economy Of A Large Area Or Population**
- **Significant Number Of Deaths**
- **Extensive National Interest, Media Coverage Or Public Scrutiny**

WHAT TYPE OF EVENTS ARE ASSESSED?

- **Any Event Which Meets The Previous Criteria, Including:**
  - **Hurricanes**
  - **Tornadoes**
  - **Floods**
  - **Winter Events**
  - **Heat Waves**
  - **Wildfires**

WHO MAKES THE DECISION?

- **The Office Of Services At National Weather Service Headquarters And The Regional Directors Of The Affected Areas.**
- **Final Approval For Assessment Comes From The NWS Director.**
- **This changed with Hurricanes Irene and Sandy (NOAA directed).**

THE SERVICE ASSESSMENT TEAM WILL NORMALLY INCLUDE:

- Subject Matter Expert For The Type Of Event Involved
- Someone With Field Experience And Current Expertise For The Event.
- Public Affairs Officer And Office Of Services Facilitator
- Someone Outside The NWS, Preferably With Expertise Related To The Event

Note… One Of These Individuals Will Serve As Team Leader

Event Specific Assessments

- **There Are Some Differences In The Assessment Approach For A Short Term Event (e.g. Tornado) vs. The Approach For A Longer Term Event (Hurricane)**

- **However… While The Assessments May Differ In Scale in Time and Space… The Goals Are Really The Same… To Evaluate What Went Right Or Wrong And Improve The Forecast and Warning Process.**

An Example…

- To Demonstrate How the Assessment Process Works… Let's Look At A Real Life Example Of What Goes Into One…

- The Following Assessment Was Conducted After A Major Tornado Event In Oklahoma City, OK in May, 2003

It Starts With Data…

- **The First Step Is To Collect As Much Data As Possible, Including:**
  - **Meteorological Data... Model Output, Radar, Satellite, Guidance Products Etc. Available *At The Time Of The Event***
  - **Products & Services… Forecasts, Warnings, Statements, Event Logs, Communications Logs, Equipment Status, Contact Information, Staffing Levels, Outreach History, Training Records… etc.**
  - **Event Data…Rough Estimate Of Timing, Areal Extent, Type of Event, Deaths, Injuries, Damage**
  - **Potential Interview List… Emergency Managers, Eyewitnesses, Media, First Responders, etc.**

After The Team Arrives On Site…Field Operations Begin

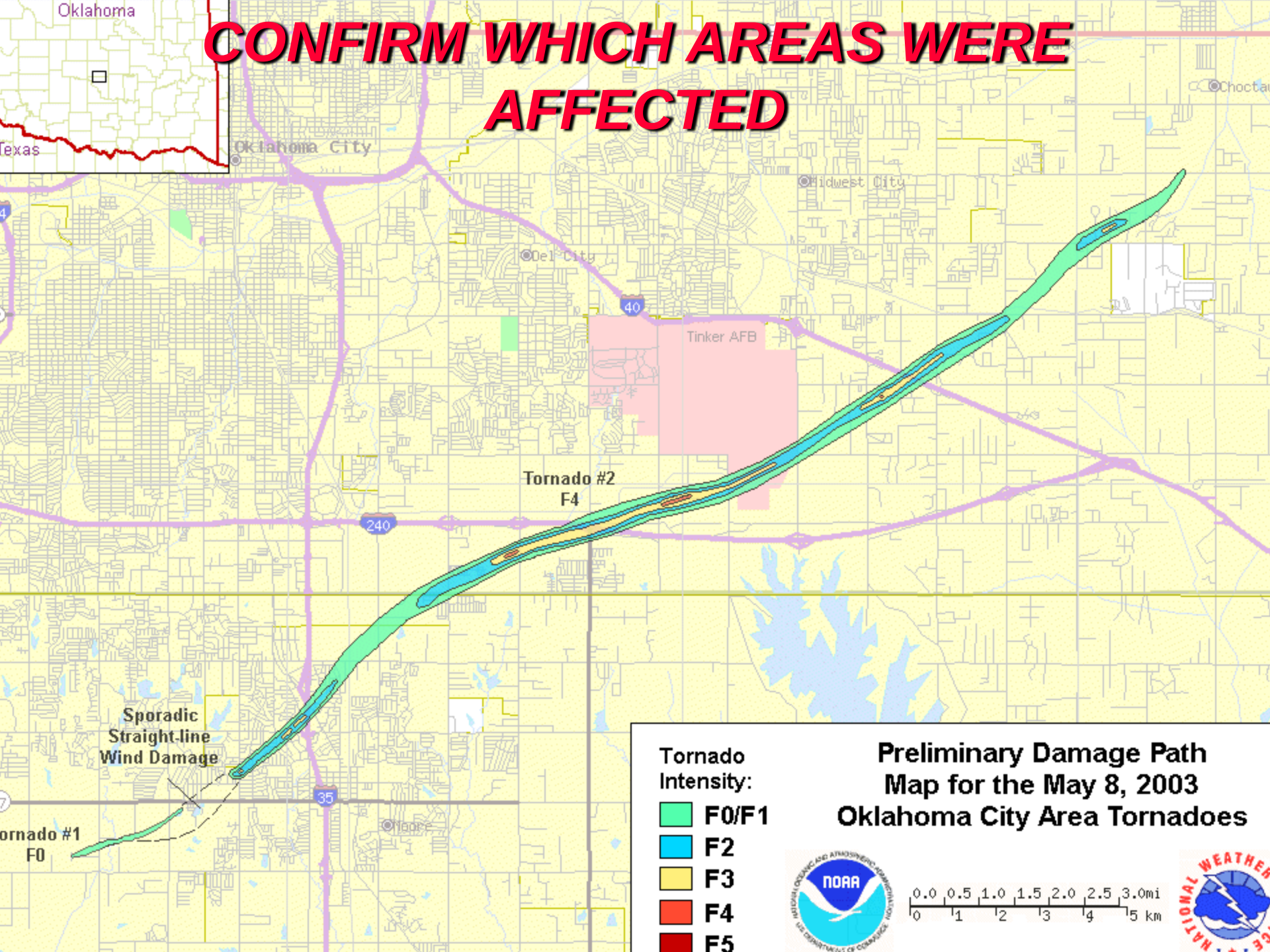
Usually, The Team Will Split Up… With Different Members Collecting Different Information

- Some Will Conduct Visual Inspections

- Some Will Interview Important Contacts

- Some Will Review Forecast Products & Performance

CONFIRM WHICH AREAS WERE AFFECTED



VISUAL INSPECTION OF AFFECTED AREAS

ESTABLISH MAGNITUDE OF DAMAGE INCLUDING NUMBER OF DEATHS, INJURIES, ECONOMIC IMPACT

The Service Assessment Team May Also Help To Establish A "Rating" For The Event

(For Example, EF-scale Rating For Tornadoes, or Saffir-Simpson Scale Rating For Hurricanes)

Normally, The Final Determination Of The Rating Will Be Made By A Quick Response Team Of Subject Matter Experts.


CONDUCTING ASSESSMENT AS QUICKLY AS POSSIBLE AFTER THE EVENT IS CRITICAL!!!

PEOPLE TO BE INTERVIEWED:

- **SURVIVORS/WITNESSES**
- **EMERGENCY MANAGERS**
- **MEDIA**
- **RESCUE PERSONNEL**

INTERVIEW TOPICS

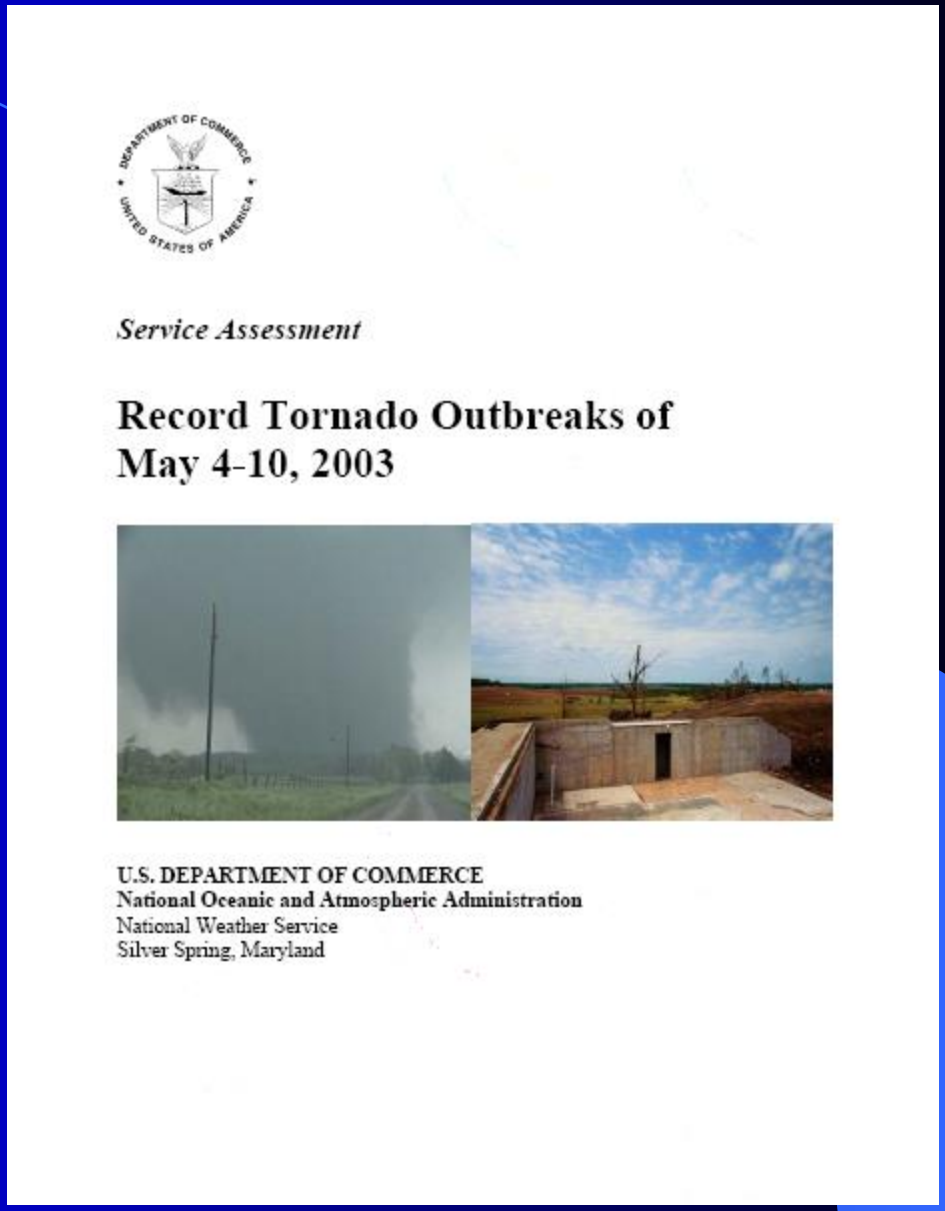
- **DID YOU RECEIVE A WARNING?**
- **WAS IT TIMELY?**
- **WAS IT UNDERSTANDABLE?**
- **HOW DID YOU RECEIVE IT?**

FORECAST OFFICE INTERVIEWS

- REVIEW WARNING PRODUCTS
- ESTABLISH LEAD TIMES
- WERE AGENCY AND OFFICE POLICIES AND PROCEDURES PROPERLY FOLLOWED?
- WHAT WENT RIGHT: “BEST PRACTICES”
- HOW CAN WE IMPROVE?

Results Of The Assessment Are Shared With All NWS Offices… So That Everyone Can Learn From Both The Positives And Negatives Of The Event

Information Sharing Of This Type Is One Of the Most Vital Methods Available To Ensure The Quality Of Our Most Important Service… Protecting Lives & Property!



Service Assessment

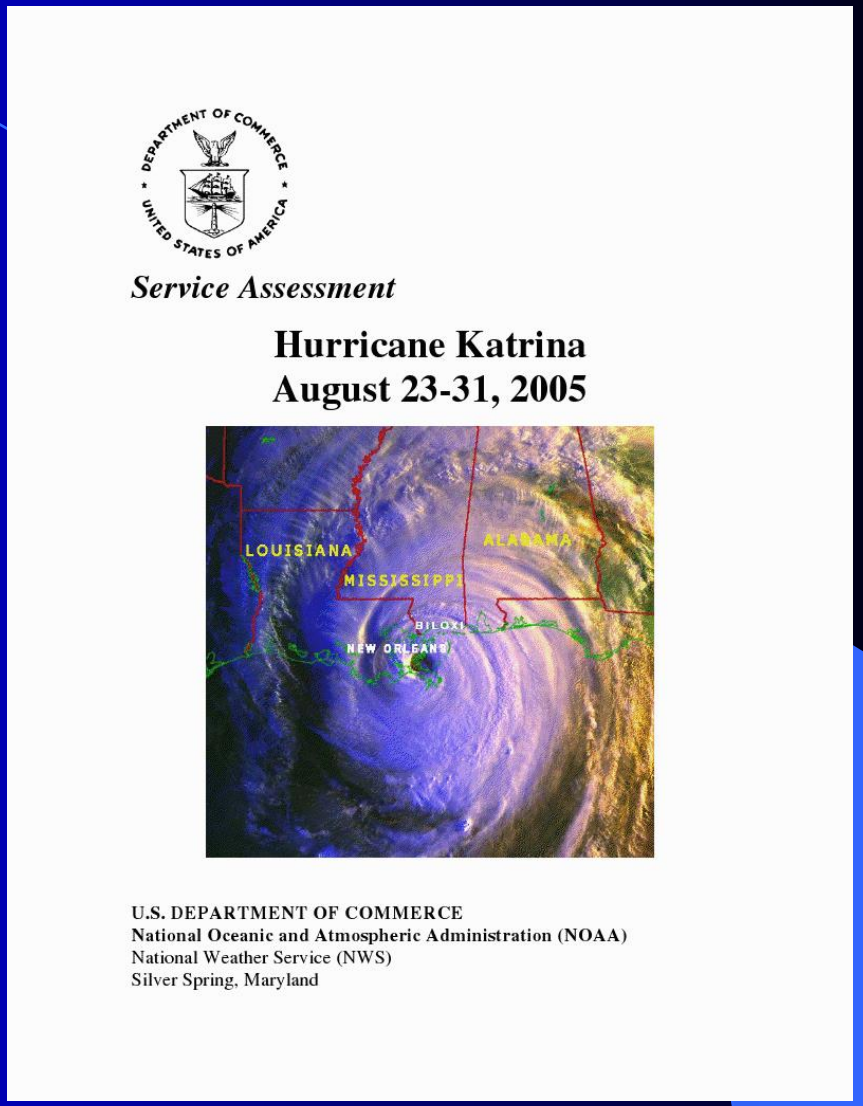
Record Tornado Outbreaks of May 4-10, 2003

U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Weather Service
Silver Spring, Maryland

So… The Assessment Identifies A Need For A New Service…

How Does That New Service Come About?

Often, The Service Assessment Report Itself Will Contain Specific Recommendations That The NWS Can Implement Through Internal Procedures



There Are Also Times When Initiatives From The Research Community Can Directly Result In A Procedural Change!

An Example:

The Enhanced Fujita Scale

The Key Points…

- Change Can Be Difficult… But If Quality Of Service Is To Improve, We Must Be Open To Change.
- We Must Constantly Be Open To The Possibility That There Are Better Ways To Do Things!
- We Do Not Know It All! Our Partners Are Often Much More Aware Of How Well Our Services Are Meeting Needs!


THE END
QUESTIONS???