5.7

# PROBABILISTIC HAZARD INFORMATION (PHI): HIGHLIGHTING THE BENEFITS VIA NEW VERIFICATION TECHNIQUES FOR FACETS

*Gregory J. Stumpf[1,2,*], Christopher D. Karstens[1,3], and Lans P. Rothfusz[3]*

*[1]Cooperative Institute for Mesoscale Meteorology Studies, Univ. of Oklahoma, Norman, OK*
*[2]NOAA/National Weather Service Meteorological Development Laboratory, Silver Spring, MD*
*[3]NOAA/National Severe Storms Laboratory, Norman, OK*

## 1. INTRODUCTION

National Weather Service (NWS) severe thunderstorm and tornado warnings are issued today as deterministic polygons that cover a two-dimensional area of expected threat of severe weather within a time period usually from 30 to 60 minutes. As part of the Forecasting A Continuum of Environmental Threats (FACETs) initiative, new methods of delivering severe thunderstorm and tornado forecast and warning information are being investigated. A concept known as Probabilistic Hazard Information (PHI) has been under development and testing at the National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) since 2008. Within the PHI concept, customized severe weather warning products can be derived from rapidly-updating probability grids that are produced by forecasters.

One advantage of PHI is that forecasters can communicate threat information prior to the thresholds that today's warnings are issued, both in time and space. For example, probabilistic information about threats that have yet to reach an intensity which would satisfy a forecaster's threshold decision to warn is made available to users who desire that information. In addition, information about a forecasted threat well downstream from the current threat location, well beyond the forward edge of any warning polygon, can also be conveyed. In each of these situations, a lower probabilistic threshold can be used to create custom warnings for users who are more at risk. In addition, the times and locations of higher confidence threats (e.g., tornadoes currently reported on the ground) can also be conveyed through the probabilistic grids.

A new gridded verification technique is used to show how warning accuracy is affected using varying probabilistic thresholds. The technique is illustrated with the 27 April 2011 tornado outbreak event over the southeast United States. Using human-generated locations of mesocyclones and tornadoes at each radar volume scan interval, swaths of strike probabilities using the Thunderstorm Environment Strike Probability Algorithm (THESPA; Dance et al. 2010) technique are generated. Innovative geospatial warning verification techniques, which include new measures such as false-alarm area and location-specific lead and departure time, will be employed to highlight the benefits of PHI-derived custom warnings.

*Corresponding author contact information: NSSL/WRDD, National Weather Center, David L. Boren Blvd., Norman, OK, 73072, greg.stumpf@noaa.gov, 405-325-6773.

## 2. LIMITATIONS OF CURRENT NWS VERIFICATION TECHNIQUE

For several decades prior to 2007, NWS severe thunderstorm and tornado warnings were issued on a county basis. County issuance was employed because most of the dissemination systems of the earlier part of that period were county based (e.g., weather radio, sirens, television graphics). Verification was also based on counties. A storm event must occur within a warned county during the duration of the warning in order for that warned county to be considered verified. If a warning was issued for multiple counties at once (e.g., for a long squall line), then each county was verified as a separate entity, and each required at least one storm event during the duration of the warning in order for that warned county to be considered verified. A multiple-county warning could only be partially verified if some of the counties were not accompanied by a storm report.

The NWS transitioned a storm-based warning system in 2007. This transition was motivated, in part, by advances in digital dissemination systems such as GPS-enable mobile devices, which would alert if the device was within the warning polygon. With the use of software known as WarnGen, NWS forecasters can now draw warnings as a polygon representing the swath that the severe weather is expected to cover within the duration of the warning, without regard to geo-political boundaries such as county lines. The storm-based warning concept was also designed to limit the warning area to that expected to be covered by the storm threat, with the intent to make warnings smaller and more precise.

Warning verification also changed with the new system. It was no longer required that each county covered in a warning received at least one storm report. Now, only one storm report within the polygon is required to verify a single storm-based warning. An advantage of this new verification system is if a storm-based warning touched several small portions of multiple counties, there was no need to verify each of those county

segments with individual storm reports. However, there are also flaws with this system that will be shown in this paper.

Forecasts are typically verified for accuracy via the use of a 2×2 contingency table (Fig. 1). Each of the four cells in the2x2 table is explained as follows: A verified forecast is defined as a *hit (A)*, and represents when an event did happen when the forecast said it would happen. An unverified forecast is defined as a *false alarm (B)* when a forecast was issued, but an event did not happen. An event that went unforecasted is defined as a *miss (C)*. And finally, wherever and whenever events did not happen, when there was a forecast of no event (or no forecast at all), that is defined as a *correct null (D)*.

| Contingency Table | | Observation | |
|---|---|---|---|
| | | Yes | No |
| Forecast | Yes | A = Hit | B = False Alarm |
| | No | C = Miss | D = Correct Null |

*Figure 1. A 2x2 contingency table.*

A number of accuracy measures can be derived from the 2x2 table. The first is called Probability Of Detection (POD), which is the ratio of verified forecasts or hits (*A*) to the number of all forecasts (*A* + *C*). Another is the False Alarm Ratio (FAR), or Probability Of False Alarm (POFA), which is the ratio of false forecasts (*B*) to all forecasts of an event (*A* + *B*). Finally, one can represent the combination of both POD and FAR into the Critical Success Index (CSI), which is the ratio of hits to the sum of all hits, misses, and false alarms. CSI can be written both as a function of *A*, *B*, and *C*, and through algebraic manipulation, a function of POD and FAR (Fig. 2).

$$\text{FAR (False Alarm Ratio)} = \frac{B}{A+B} \qquad \text{POD (Prob. Of Detection)} = \frac{A}{A+C}$$

$$\text{CSI (Critical Success Index)} = f(\text{FAR, POD}) = \frac{1}{\frac{1}{1-\text{FAR}}+\left(\frac{1}{\text{POD}}-1\right)} = \frac{A}{A+B+C}$$

*Figure 2. Equations for POD, FAR, and CSI given the 2x2 table in Fig. 1.*

For the NWS storm-based warning verification system, storm report *points* are used to verify warning *areas*. If there is a severe weather report point falling outside any warning polygons, that report point is counted as one *miss*. If a warning area (the polygon) contains no severe weather report points, the warning area is considered one *false alarm*.

Hits are counted for both report points and warning areas. If a severe weather report point is inside a warning polygon, then that report point is counted as one *hit*. If a warning area (the polygon) contains at least one report point, that warning area is considered a hit. Both hit values are used for NWS verification.

Given this, we must consider two 2x2 contingency tables for verifying storm-based warnings, one table for the warning areas or polygons (Fig. 3) and one table for the report points (Fig 4.). To compute false alarm ratio, the NWS uses the 2x2 table for warning areas. We will refer to the accuracy measures from this 2x2 table with the subscript of 1.

| Contingency Table 1 | | At Least One Storm Report Point at Any Time within Polygon | |
|---|---|---|---|
| | | Yes | No |
| Warning Polygon Exists | Yes | $A_1$ = Hit | $B_1$ = False Alarm |
| | No | $C_1$ = Miss | $D_1$ = Correct Null |

*Figure 3.The 2x2 contingency table for report points.*

$$\text{FAR}_1 = B_1 / (A_1 + B_1) \qquad (1)$$

To compute probability of detection, the NWS uses the 2x2 table for report points. We will refer to the accuracy measures from this 2x2 table with the subscript of 2.

| Contingency Table 2 | | Storm Report Point Exists | |
|---|---|---|---|
| | | Yes | No |
| Point is Within Warning Polygon | Yes | $A_2$ = Hit | $B_2$ = False Alarm |
| | No | $C_2$ = Miss | $D_2$ = Correct Null |

*Figure 4. The 2x2 contingency table for warning polygon areas.*

$$\text{POD}_2 = A_2 / (A_2 + C_2) \qquad (2)$$

The values for false alarm ($B_1$) and miss ($C_2$) are from two different tables. Also, two values of hit are used, 1) the number of verified warning areas ($A_1$), and 2) the number of verified report points ($A_2$). Because multiple reports can be contained within a single warning, $A_2 \geq A_1$.

The NWS formula for CSI uses FAR from the first 2x2 table (Fig. 3) and POD from the second 2x2 table (Fig. 4).

$$\text{CSI}^* = f(\text{FAR}_1, \text{POD}_2) = \frac{1}{\frac{1}{1-\text{FAR}_1}+\left(\frac{1}{\text{POD}_2}-1\right)} \qquad (3)$$

Substituting for FAR and POD in the above equation gives

$$CSI^* = \frac{A_1 A_2}{A_1 A_2 + A_2 B_1 + A_1 C_2} \quad (4)$$

which is not equivalent to the last equation in Fig. 2. This dilemma is addressed by the new verification scheme in this paper.

As stated above, severe storm reports are given as points, or as in the case of tornado tracks and the occasional hail track, given as a line or series of points at one-minute intervals. However, severe weather affects two-dimensional areas. Hail falls in "swaths" with width and length. Tornadoes follow paths with width and length (although usually the width is below the precision of a warning polygon, usually < 1 km). And wind damage can occur over broad areas within a storm. Area forecasts are better verified using areal data. However, the observational data used to verify areal warnings is usually lacking in completeness.

Because only one storm report verification point is required to verify a warning polygon area, a single point can be used to verify a polygon of any size or duration. As shown in Figure 5, each of these two polygon areas would be scored as hits, the small warning and the large warning. A larger in area and longer in duration warning provides greater likelihood of capturing a severe weather report within the warning area and time. There is also a greater likelihood of having *multiple* storm reports within the warning, resulting in multiple point hits ($A_2$). Because the NWS formula for POD is based on report points ($POD_2$), a forecaster issuing larger and longer warnings should inevitably be rewarded with a larger average POD of all their reports.

*Figure 5. Examples of two warning polygons, with the one of the right larger than the one on the left.*

The False Alarm Ratio (FAR) is calculated using the 2×2 contingency table for polygon areas ($FAR_1$). If any warning area ends up not verifying (no report points within the area), then that warning area gets counted as a single false alarm, *regardless* of the size and duration of that warning area. For a forecaster, there is no false alarm penalty for making warnings larger and longer, and as shown earlier, improves the chances of a higher POD by potentially capturing more severe weather points. Furthermore, if the warning is issued with a very long duration and well before anticipated severe weather, this increases the chances of having a larger lead time. If that warning never verifies, there is no negative consequence on the calculation of average lead time for all warned events, because lead time requires a severe weather report to be computed. "Casting a wide net" has its advantages – it potentially improves POD and lead time without significantly impacting FAR. This ends up being counter to the stated benefits of storm-based warnings – the reduction in the size of warnings and greater warning precision and specificity. Note too that for large and long warnings, even if they verify, there can a tremendous amount of area falsely warned and for a large amount of time.

We address these pitfalls with an improved warning verification methodology that will better reward precision in warnings, and provide a host of other benefits, including ways to measure the goodness in new warning services techniques.

## 3. THE NEW VERIFICATION TECHNIQUE

### *3.1 The Grids*

To consolidate the verification measures into one 2×2 contingency table and reconcile the area versus point issues, we place the forecast and observation data into the same coordinate system. This is facilitated by using a gridded approach, with a spatial and temporal resolution fine enough to capture the detail of the observational data. For this study, we use a grid resolution of 1 $km^2$ × 1 minute, which should be fine enough to capture the events with the smallest dimensions.

The *forecast grid* is created by digitizing the warning polygons (defined as a series of latitude/longitude pairs), so that points inside (outside) the polygon are assigned a value of 1 (0). The grid has a 1 minute interval, so that warnings appear (disappear) on the grid the exact minute they are issued (canceled/expired). Warnings polygon coordinates that are also modified via use of Severe Weather Statements (SVS) are reflected as changes in the forecast grid.

The *observation grid* can be created using either ground truth information [e.g., NWS Local Storm Reports (LSR)], ground truth data augmented by human-determined locations using radar and other data as guidance, or radar proxies, or a combination of all of these. Line data in the case of tornado and hail tracks are divided into one-minute increments, with each increment treated as an observation point. For the purposes of basic NWS warning verification, the point data can be converted to a truth grid by applying a sphere of influence or "splat" around the data point. The observation splat is used to, 1) account for small uncertainties in the timing and position of the report locations, and 2) to account for an average "safety

zone" around events that might be derived from user opinion surveys on how close one should be to an event to warrant a warning. The splat distance can be set to the grid resolution (e.g., 1 km) if no splat is desired.

There is one more optional grid that can be used for computing the scores. When determining a correct null (*d*) forecast, using every grid point outside of each warning polygon and each observation splat, the value of *d* would overwhelm all other grid points; tornadoes, and even tornado warnings, are rare events (Marzban, 1998). We can limit *d* by excluding those points where it is obvious that a warning should not be issued – namely grid points which are outside of storm areas. This can be done by thresholding by radar reflectivity or any other quantity that can be used to define storm locations. The grid points outside of a warning polygon, an observation splat, and storm areas, make up the *exclusion grid*, and would not be used to calculate the accuracy measures. Given the forecast grid, the observation grid, and the optional exclusion grid, there are two methods for measuring accuracy.

### *3.2 Grid Point Scoring Method*

We can use the above grids in a single 2x2 contingency table (Fig. 6). At each one-minute forecast interval, we can create these grid point values, shown graphically for a hypothetical storm, warnings, and report point "splat" area in Fig. 7:

HIT: The grid point is warned AND the grid point is within the splat range of a report point (*A*).

FALSE: The grid point is warned AND the grid point is outside the splat range of any report point (*B*).

MISS: The grid point is not warned AND the grid point is within the splat range of a report point (*C*).

CORRECT NULL: All other grid points OR all other grid points not within the exclusion grid, if used (*D*).

NON-EVENT: All grid points within the exclusion grid, if used (variable not used).

Accuracy measures from the 2x2 table for each grid point, at each time step, can be computed. These include POD, FAR, a CSI derived from this single 2x2 table, and other measures that include *D* such as Heidke Skill Statistic (HSS).

| Contingency Table | | Grid point within range of tornado hazard at that time | |
|---|---|---|---|
| | | Yes | No |
| Grid point is warned at that time | Yes | A = Hit | B = False Alarm |
| | No | C = Miss | D = Correct Null |

*Figure 6. 2x2 contingency table for grid points.*

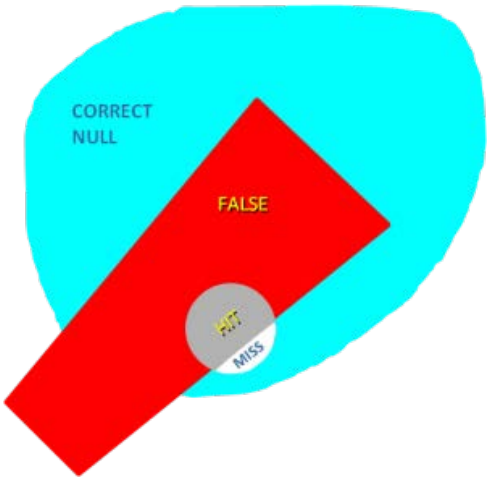


*Figure 7. Hypothetical storm, warning, and report point "splat" area. The correct null area (cyan) roughly outlines the "storm"; the warning polygon is comprised of the false alarm area (red) and the hit area (grey); the report point area is comprised of the miss area (white) and shares the hit area (grey) with the warning polygon. All grid points outside of these areas are considered non-events.*

### *3.3 Truth Event Scoring Method*

Additional scoring methodologies are needed to address the issue of lead time (LT) at specific grid point locations. One must consider grid points that are downstream of an observation event that are expected to be covered by that event on a future grid. In addition, grid points that are behind an event that has already passed that location can be considered for a new metric, departure time (DT). This measures the amount of time a location remains under a warning after the threat as passed, which should be minimized.

To build "truth events", we look at the timeline of the Grid Point Method score for specific grid point locations. As storm events and warnings pass over specific grid points, the score conditions for those grid points will vary between non-event, correct null, false, miss, and hit. A "truth event" is defined as a continuous time period a specific grid point is under a warning(s) and/or a storm observation(s) and/or surrounded by at least one minute of a non-event condition (none of the four conditions listed below). The following types of truth events can be recorded:

FALSE EVENT: If grid point remains in "false" condition throughout event (only forecast grid becomes "1"). These grid points do not receive an observation of a hazard, but were warned.

MISS EVENT: If grid point remains in "miss condition" throughout event (only observation grid becomes "1"). These grid points were not warned, but received an observation of a hazard.

HIT EVENT: If grid point experiences a "hit condition" for at least 1 minute during event (forecast grid and observation grid are both "1"). These grid points were warned AND received an observation of a hazard.

CORRECT NULL EVENT: If grid point remains in "correct null condition" throughout event (neither the observation grid nor the forecast grid becomes "1"). These grid points were not warned, and did not receive an observation of a hazard.

Hit Events can then be comprised of several different scenarios. The most common scenario would be this: 1) a warning is issued for a particular grid point, 2) a hazard observation impacts that grid point after warning issuance, 3) the hazard observation leaves the grid point while the warning remains in effect, and 4) the warning ends for that grid point (via expiration or cancellation). For these scenarios, the grid points will be in false condition prior to and after the hazard passes over that location. For the Truth Event method, these conditions are not considered false, but instead are depicted as *lead time* and *departure time*, respectively.

For this common scenario, the truth event is defined by starting and ending time of the warning. Since the warning was issued prior to the observation impacting the grid point, there is positive lead time. If an observation impacts a grid point prior to a warning, then we measure *negative* lead time. If an observation impacts a grid point that is never warned (a Miss Event), then no – *not zero* – lead time is recorded. This differs from the current NWS verification method, which records a zero (0) minute lead time for missed events. In essence, missed events are treated the same as warned events that were warned "just in time", which is clearly incorrect.

This verification method also allows us to analyze a new kind of metric called *departure time*. This is the amount of time that a grid point remains under a warning after the threat has already passed. Ideally, the departure time should be zero – the warning is canceled or expires immediately just after the threat has passed. Positive departure time, as with the first scenario, is chosen to represent the condition when the warning remains in effect after the threat has passed (a false condition, in a sense). Negative departure time is chosen to represent the condition when the warning has expired or was canceled before the threat has passed (a miss condition, in a sense).

We can also analyze a third kind of metric which called *false alarm time* (FAT). This is for False Events - events that remain in false condition through their time period.

For each truth event, the following quantities are calculated:

LEAD TIME (LT): $t_{obsBegins} - t_{warningBegins}$

DEPARTURE TIME (DT): $t_{warningEnds} - t_{obsEnds}$

FALSE ALARM TIME (FAT): $t_{warningEnds} - t_{warningBegins}$

where $t_{warningBegins}$ = time that the warning begins, $t_{warningEnds}$= time that the warning ends, $t_{obsBegins}$ = time that the observation begins, and $t_{obsEnds}$= time that the observation ends. LT and DT are only calculated for Hit Events. FAT is only calculated for False Events.

The number of grid points (1 $km^2$) that are warned falsely is used to calculate an additional metric, the False Alarm Area (FAA). Both FAT and FAA are important measures which reward precision in warnings. In other words, warnings that are made small enough in time and space as to capture the event but not be too large to overwarn users away from the threat.

## 4. DATA

The verification technique is tested on data collected during the 27-28 April 2011 super-outbreak of tornadoes. Two data sets are used, and both are scored for Tornado Warnings against tornado observations.

The first data set uses only the single long-tracked tornadic supercell that affected Tuscaloosa and Birmingham Alabama during the afternoon and evening of 27 April 2011 (hereafter, the "TCL storm"). Figure 8 shows composited radar images of the storm that moves across the entire state of Alabama from southwest to northeast. This particular supercell produced two long-tracked violent tornadoes within Alabama (Fig. 9). The domain area is comprised of the Birmingham, Alabama, NWS Weather Forecast Office (BMX WFO) county warning area (CWA).

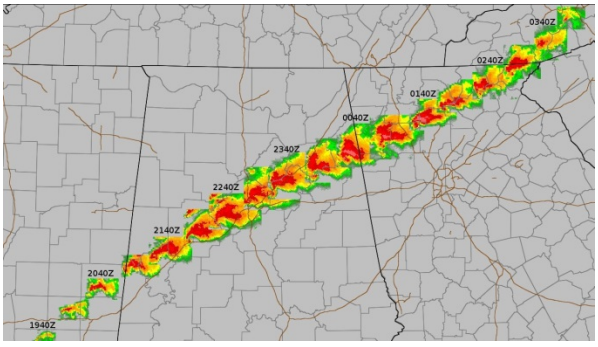


*Figure 8. Radar images of the TCL storm composited over the lifetime of the event. Times in UTC from 27-28 April 2011 are annotated. Image courtesy NCAR.*

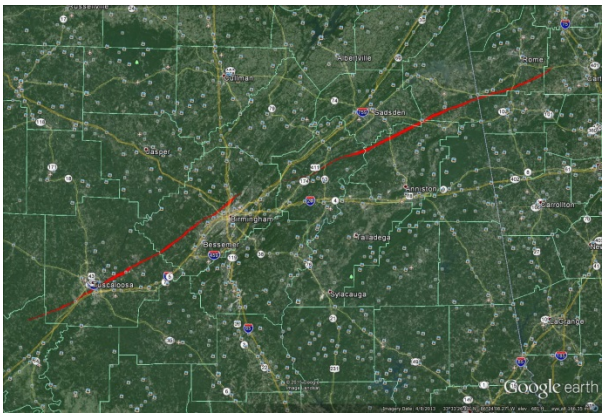


*Figure 9. Damage paths (red) of the two violent tornadoes associated with the TCL storm. Shapefile courtesy of the NWS Damage Assessment Tool.*

The second data set is comprised of every tornadic and non-tornadic supercell from the outbreak during the afternoon and overnight period from 1830 UTC 27 April 2011 – 0900 UTC 28 April 2011 (Fig. 10). The domain area is comprised of the CWA from these four WFOs: Jackson, Mississippi (JAN), Birmingham, Alabama (BMX), Huntsville, Alabama (HUN), and Peachtree City, Georgia (FFC).

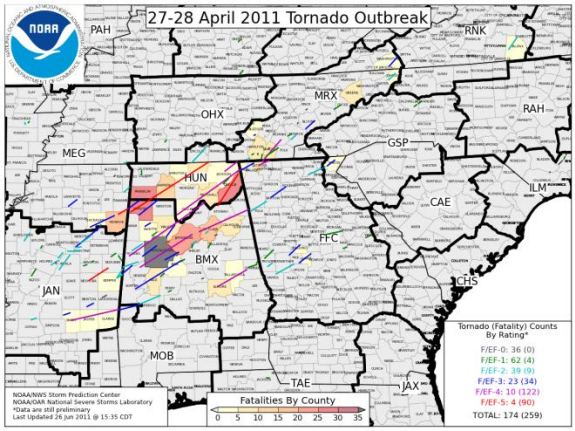


*Figure 10. Damage paths of all tornadoes during the 27-28 April 2011 outbreak. Image courtesy NWS Service Assessment (NWS, 2011).*

The observation grid is created by "truthing" the centroid locations of both tornadoes using radar data and damage survey information. The tornado track information does not have the precise temporal information along the path of the tornado, so radar data was used to determine the location of the tornadoes with respect to time, roughly at 4-5 minute intervals. The tornado locations were then interpolated at precisely one-minute intervals (e.g., 00:01:00, 00:02:00, in hh:mm:ss). These locations were used to create 1 $km^2$ grids with a "splat" radius of 5 km.

An exclusion grid for determining correct null information is created using a median-filtered composite reflectivity (CREF; maximum reflectivity in the vertical column) field from the Multiple-Radar / Multiple-Sensor (MRMS) system (Lakshmanan et al, 2006). Grid points where CREF < 30 dB*Z* at each 1 minute interval are excluded from processing.

Three types of forecast grids are used to compare warning accuracy.

### *4.1 Official NWS Warnings*

For the first data set, the BMX WFO issued 6 separate Tornado Warnings covering the TCL storm over 4 hours and 22 minutes (2038 UTC 27 April 2011 – 0100 UTC 28 April 2011), from the AL-MS border northeastward to the AL-GA border. Each of the warnings was modified during their durations follow-up Severe Weather Statements (SVS) in which the polygons were reduced in size by forecasters by manually removing warning areas behind the threats. For the second data set, the JAN, BMX, HUN, and FFC WFOs issued more than 200 Tornado Warnings over the course of the data set (1830 UTC 27 April 2011 – 0900 UTC 28 April 2011) (Figure 11).

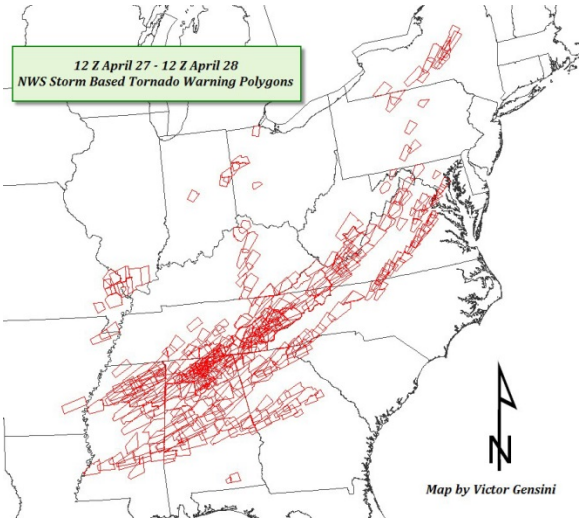


*Figure 11. All Tornado warnings for the 27-28 April 2011 outbreak. Map courtesy Victor Gensini.*

### *4.2 Threats-In-Motion*

With the "Threats-In-Motion" (TIM) concept, the warning polygons essentially follow the storm. The leading edge of each polygon would inch downstream with the threat, and the trailing edge would automatically clear from areas behind the threat. We hypothesize that this would result in a larger average lead time for users downstream, which is desired. In addition, the departure time of the warning should approach zero, which is also considered ideal. The TIM concept is essential for any future hazard warning system that is

based on probabilistic information, because probabilities evolve continuously across time and space.

In order to create a set of TIM warnings, a second observation grid is used. This grid is created by "truthing" the centroid locations of the human-inferred locations of radar-based mesocyclones and tornadic vortex signatures (TVS) during the history of the storm. Where mesocyclones/TVS overlap reported tornadoes, the locations of the tornadoes are substituted along the track. As before, the mesocyclone/TVS locations were then interpolated at precisely one-minute intervals. These locations were used to create warning polygons using the default AWIPS WarnGen polygon shape parameters. The current threat point is projected to its final position using the motion vector and a prescribed warning duration – 45 minutes because most of the official NWS warnings for the storms were around 45 minutes. A 10 km buffer is drawn around the starting threat point resulting in the back edge of the warning being 20km. A 15 km buffer is drawn around the ending threat point resulting in the back edge of the warning being 30km. The far corners of each box are then connected to create the trapezoid shape of the default warning polygon (Fig. 12). A new warning polygon is redrawn at every one-minute interval, resulting in a threat polygon that is continuously "in motion".

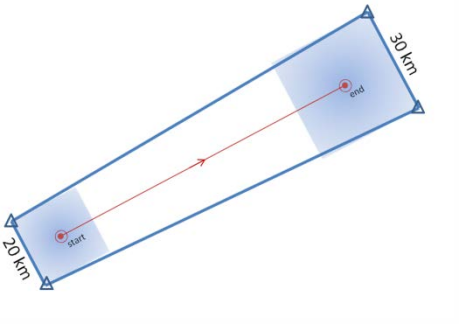


*Figure 12. The default warning polygon that is produced by AWIPS WarnGen.*

*4.3 Probabilistic Hazard Information (PHI)*

As part of the Forecasting A Continuum of Environmental Threats (FACETs) initiative, a concept known as Probabilistic Hazard Information (PHI) has been under development and testing at the National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) since 2008. Within the PHI concept, customized severe weather warning products can be derived from rapidly-updating probability grids that are produced by forecasters (Karstens et al, 2015). Given a probabilistic grid, "warnings" can be derived as an outline of a particular probabilistic threshold value. Low (high) probability warnings can be used for users who are more(less) vulnerable to hazards. Middle of the road probabilities could be used to derive "legacy" warnings based on today's forecaster uncertainty (a value that would need to be determined via social science).

The THunderstorm Environmental Strike Probability Algorithm (THESPA, Dance et al. 2010) is used to create probabilistic swaths for the mesocyclone centroid locations in each data set (Fig. 13). THESPA swaths are based on a statistical strike probability model that was developed using a database of storm detection centroids. The model determines probabilities where the storm areas will affect within a certain time period. It does not attempt to determine probabilities that a particular storm is severe or tornadic. Because probabilities evolve with time, it makes little sense to issue probabilistic swaths at regular warning intervals of 30-, 45-, or 60-minutes. Otherwise, there would be dramatic changes in probabilities at downstream locations with each update. Thus, the probability swaths are updated every minute in the same fashion as TIM. Using the low (high) probability values as thresholds for warnings result in warning areas that are larger (smaller) than middle probability values (Fig. 14). Because the low (high) probability swaths will impact downstream locations earlier (later) than a middle probability swath, the result will be that impacted locations will be warned with potentially longer (shorter) lead time than middle probability values.

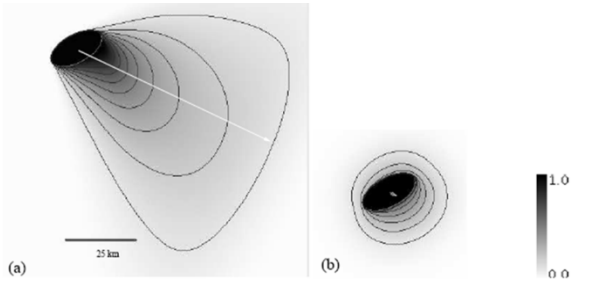


*Figure 13. Probabilistic swaths derived from THESPA for a (a) fast moving storm and a (b) slow moving storm. From Dance et al. (2010).*

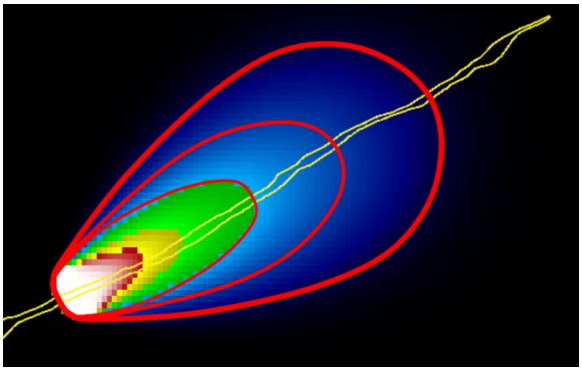

*Figure 14. THESPA-derived probabilistic grid at one time during the TCL storm. The thin yellow contour is the outline of the tornado damage survey. The red contours are three arbitrary warning "polygons", thick, normal, and thin, corresponding to low, middle, and high probability values, respectively.*

## 5. RESULTS

### *5.1 Comparing NWS warnings with Threats-In-Motion*

#### a. THE TCL STORM

All six of the NWS warnings verified in the traditional sense – each contained a report of a tornado, and both tornadoes were warned. There were no misses and no false alarms. Thus, POD = 1.0, FAR = 0.0, and CSI = 1.0. The statistics on each of the one-minute segments of the tornado paths show that 184 of 186 minutes of tornado were warned, giving a Percent Event Warned (PEW) of 0.99 (there was a 2-minute gap without a warning in effect while the storm was southwest of Tuscaloosa. For each of those 1-minute tornado segments, the average lead time was 22.1 minutes. These numbers are considered very respectable and well above the average NWS standards.

The grid-based verification method scores for the TCL event are POD = 0.9885, FAR = 0.9915, and CSI = 0.0085. The POD is very similar to the PEW computed using the NWS method. But the FAR and CSI are much different. The FAR is very large, and thus the CSI is very small. For each the accumulated one minute intervals of the 4 hours and 22 minutes of warning for this event, over 99% of the grid points within the warning polygons were not within 5 km of the tornado at each one minute interval. Because the FAR here is an accumulated measure of the amount of area (by number of grid points) and time (by number of grid intervals) that an area is falsely warned, this number can be used to determine False Alarm Area (FAA) and False Alarm Time (FAT).

Threats-In-Motion is hypothesized to improve lead time, reduce departure time, and reduce false alarm time. How does this hold up for the TCL storm? Figure 15 shows those measures for the NWS warnings and TIM warnings using the Truth Event scoring method for all 1 $km^2$ grid points within the BMX CWA at every 1-min time step over the time period of the event. All of these numbers point to a remarkable improvement using the Threats-In-Motion concept of translating the warning polygons with the storm, a truly *storm-based* warning paradigm.

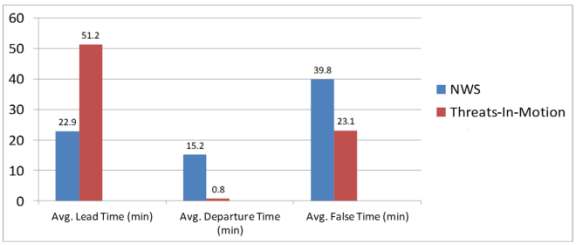


*Figure 15. Comparison of Truth Event scores for the NWS Warnings (blue) and the TIM warnings (red) for the TCL storm.*

The average values of lead time for all grid points impacted by the tornado (plus the 5 km "splat") are *more than doubled* for the TIM warnings (51.2 minutes versus 22.9 minutes). Figure 16 shows the location-specific lead time distribution with time on histograms. For the TIM warnings, there are a lot more values of Lead Time above 40 minutes. Figure 17 illustrates the geospatial distribution of lead time values. There are sharp discontinuities of lead time (from nearly zero minutes to over 40 minutes) at the downstream edges of the NWS warnings (indicated by yellow arrows in the top part of Fig. 17). These discontinuities are virtually eliminated with the TIM warnings. There are a few remaining discontinuities with TIM; these small differences are caused by changes in the storm motions at those warned times.

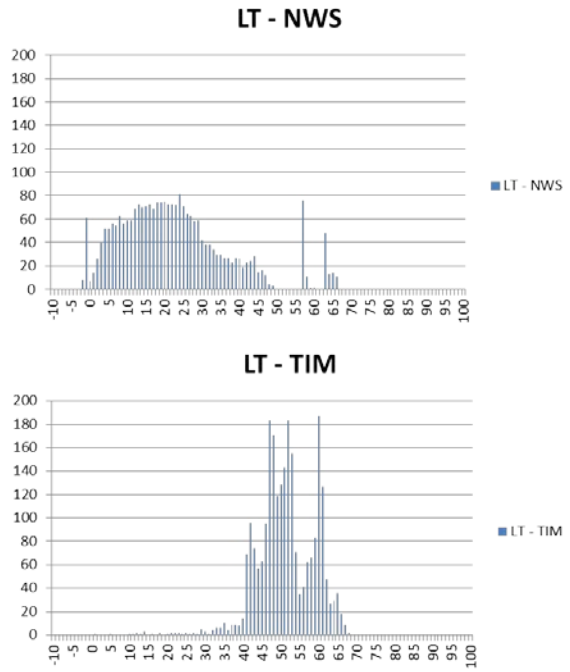


*Figure 16. Histograms of location-specific lead time. NWS warnings on the top, TIM warnings on the bottom.*

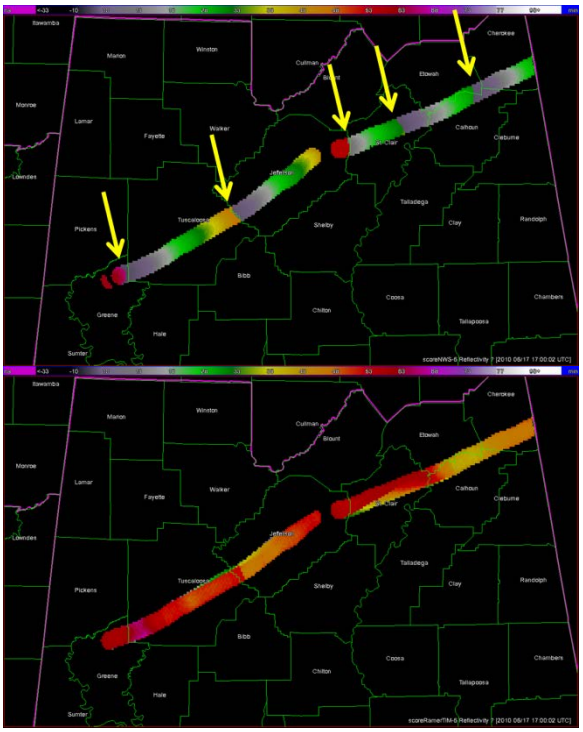

*Figure 17. Location-specific lead times to the two tornadoes for the TCL storm. NWS warnings are at the top, with lead time discontinuities indicated with yellow arrows. TIM warnings are at the bottom.*

The average value of departure time for all grid points within the splat zone of the tornado are *reduced to nearly zero* for the TIM warnings (0.8 minutes versus 15.2 minutes). Figure 18 shows the location-specific departure time distribution with time on histograms. With the TIM polygons, the departure times across the path length of both tornadoes is pretty much less than 3 minutes everywhere. Whereas, the NWS polygon Departure Times are much greater, and there are some areas still under the NWS warning more than 30 minutes after the threat had passed. Figure 19 illustrates the geospatial distribution of departure time values.

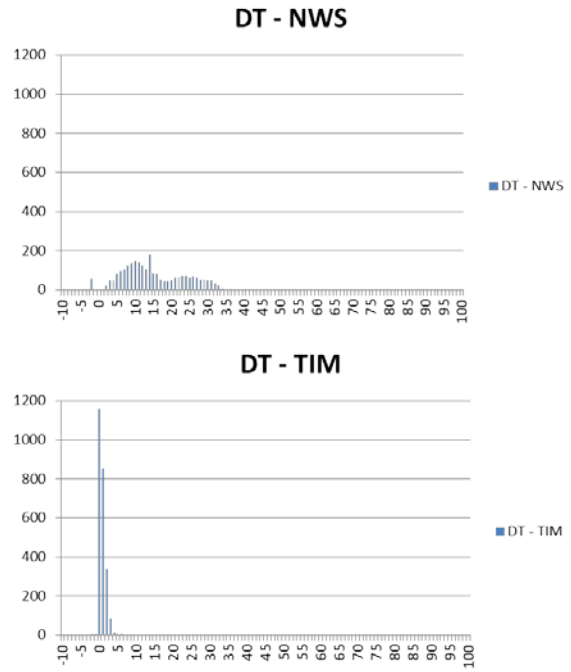


*Figure 18. Histograms of location-specific departure time. NWS warnings on the top, TIM warnings on the bottom.*

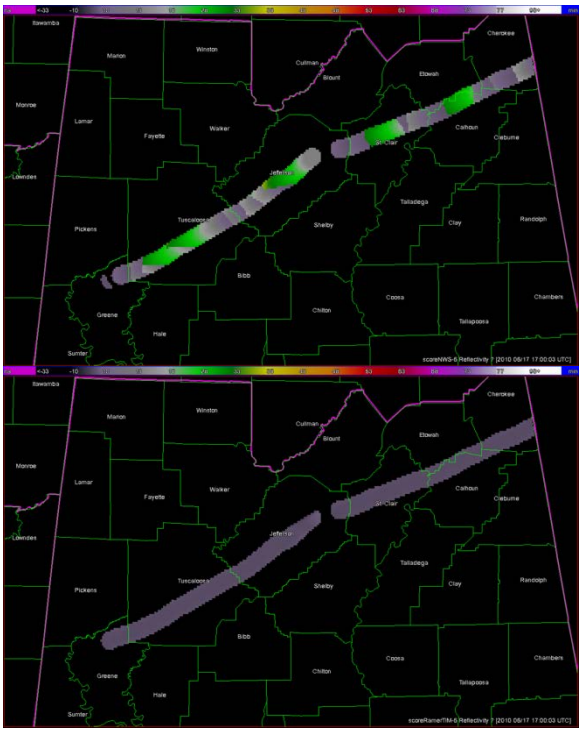

*Figure 19. Location-specific departure times to the two tornadoes for the TCL storm. NWS warnings are at the top. TIM warnings are at the bottom.*

The average value of false alarm time (FAT) for all grid points within the splat zone of the tornado is *cut nearly in half* for the TIM warnings (23.1 minutes versus 39.8 minutes). Figure 20 shows the location-specific false alarm time distribution with time on histograms. Figure 21 illustrates the geospatial distribution of false alarm time values. There are some large areas within the NWS polygons that are under false alarm for over 50 minutes at a time, even though these warnings would have verified perfectly in the traditional sense. In comparison, the TIM warnings have a much smaller average false alarm times for areas outside the tornado path (about a 42% reduction).

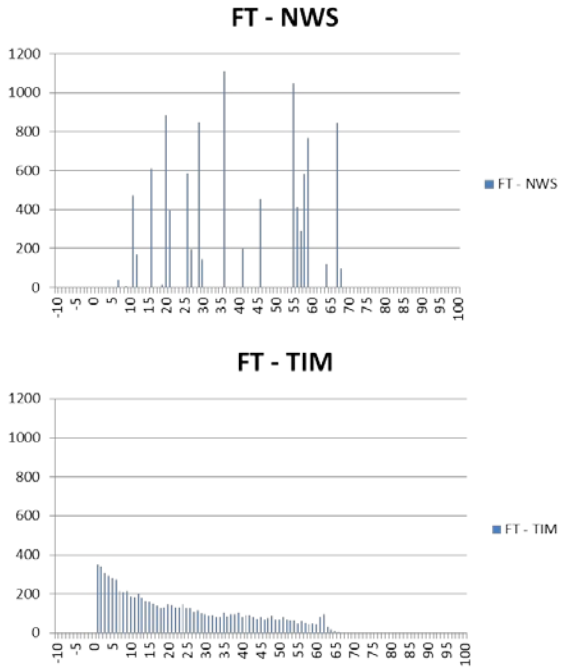


*Figure 20. Histograms of location-specific false alarm time. NWS warnings on the top, TIM warnings on the bottom.*

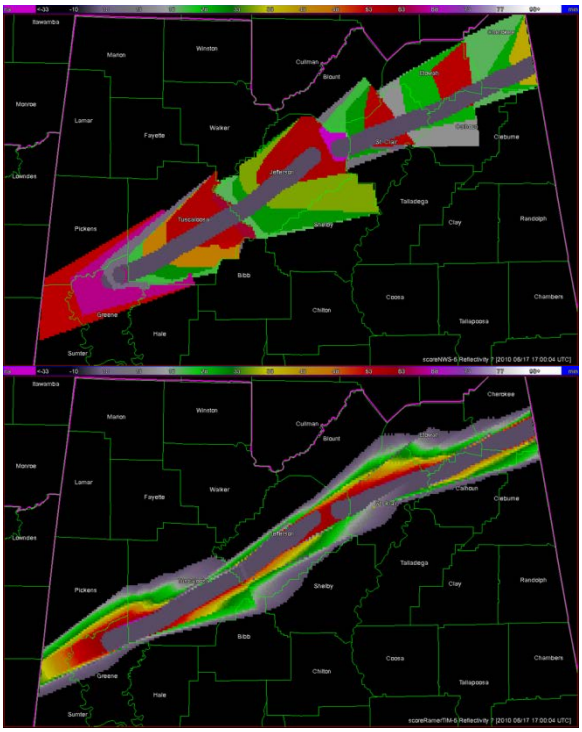

*Figure 21. Location-specific false alarm times to the two tornadoes for the TCL storm. NWS warnings are at the top. TIM warnings are at the bottom.*

The False Alarm Area (FAA) for the NWS warnings is 10,304 km$^2$ and for the TIM warnings, FAA is 8,103 km$^2$. That is a 21% reduction in FAA with the TIM warnings. The reduction in FAA is a function in the size of the warning polygons. The WarnGen defaults were used for our TIM polygons, but it appears that the NWS polygons were made a little larger than the defaults for this event.

### b. THE AFTERNOON/OVERNIGHT TORNADO OUTBREAK

The Truth Event scoring method measures for the NWS warnings and TIM warnings for the entire afternoon and overnight tornado outbreak from 1830 UTC 27 April 2011 – 0900 UTC 28 April 2011 over the domain comprised of the JAN, BNX, HUN, and FFC CWAs are shown in Figure 22. These numbers show similar improvements for each measure as with just the TCL storm. One exception is the average departure times for the TIM warnings. They are not nearly zero as with the TCL storm. This is most likely due to the fact that there are training storms with overlapping TIM warnings. Beyond the scope of this paper is the need to pair storm events with warning events perhaps with an event ID in order to isolate the measures on a storm-by-storm basis. This would also help with other types of overlapping events such as storm clusters, mergers, and splits.

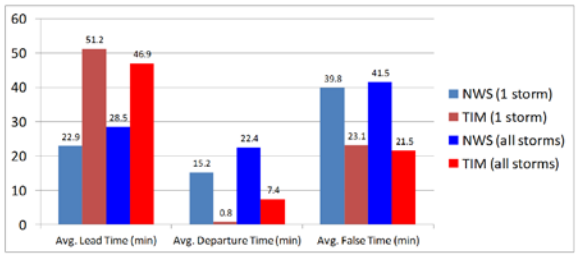


*Figure 22. Comparison of Truth Event scores for the NWS Warnings (blue) and the TIM warnings (red) for the expanded data set comprised of the TCL storm (dull colors) and all afternoon/overnight storms in the JAN-BMX-HUM-FFC domain (bright colors).*

*5.2 The effect of variable probability thresholds using PHI.*

a. THE TCL STORM

Truth Event scoring is employed on the THESPA-generated probability swaths for each one-minute interval following the TCL storm in order to determine the effect of the larger (smaller) warned areas with low (high) probability thresholds. Figure 23 depicts the variation in False Alarm Rate (FAR) and Lead Time (LT) for all truth events within the TCL storm time and space domain. As hypothesized, there is a tradeoff when decreasing probabilities to increase the warning area and length to improve lead time – the FAR also increases.

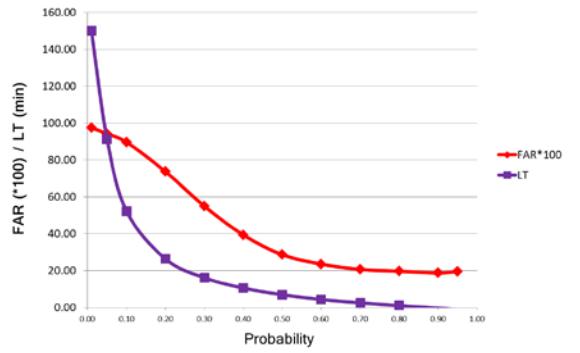


*Figure 23. Variation of lead time (purple) and false alarm ratio (red) for different probabilistic warning thresholds.*

Probabilistic forecasts can also be analyzed for reliability, which measures how close the forecast probabilities are to the true probabilities, given that forecast. Perfect reliability is achieved when the observed frequency of an event matched the forecasted probability of the event. In reliability diagrams, perfect reliability is along a diagonal from the origin to the upper rightmost corner of the diagram (100% observed frequency for a probabilistic forecast value of 1). Figure 24 depicts a reliability diagram showing the observed frequency of a forecasted event, or tornado (and the 5 km "splat") for each of 100 bins of forecast probability, from 0 to 1 for every 0.01. The forecasts are nearly perfectly reliable from 0 to about 0.7 probability. Above 0.7, the observed frequencies of the events are less than the forecasted probability, which indicates overforecasted events in those probability ranges. This is due to the fact that the tornado does not exist for the entire lifetime of the supercell storm and associated mesocyclones which were used to create the probabilistic threat in motion swaths.

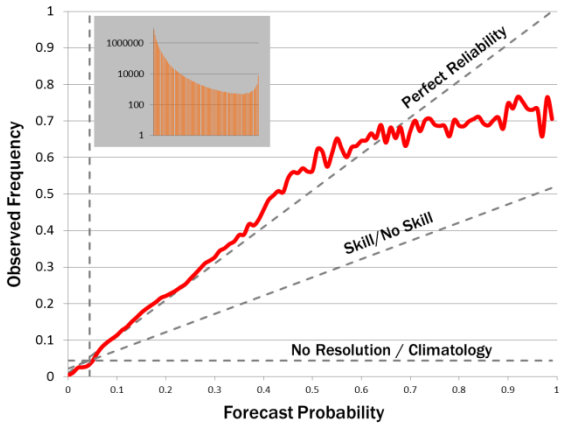


*Figure 24. Reliability diagram for all THESPA probabilities for the tornadoes with the TCL storm. The probability distribution is shown in the inset.*

b. THE AFTERNOON/OVERNIGHT TORNADO OUTBREAK

Figure 25 depicts a reliability diagram for the entire afternoon/overnight tornado outbreak data set for forecasted tornado events. The diagram shows that the forecasts are not very reliable, overforecasting across the spectrum of all probability forecasts. In addition, the forecasts show no skill - the skill/no-skill line is positioned halfway between the no resolution line (climatology) and perfect reliability. This is due to the fact that the probability swaths are generated for all mesocyclones/TVSs for the event, many of which were not associated with a tornado at every one-minute time step.

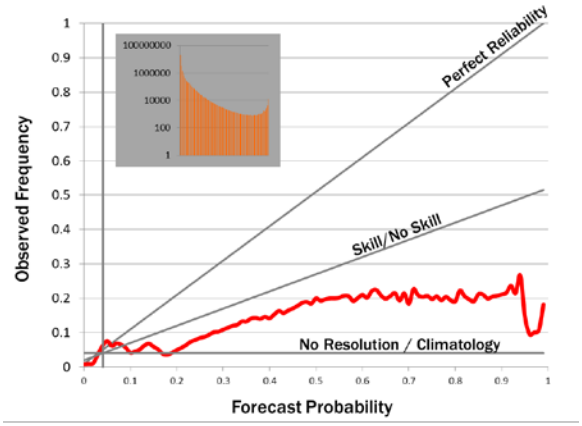


*Figure 25. Reliability diagram for all THESPA probabilities for the tornadoes with all the storms in the afternoon-overnight period. The probability distribution is shown in the inset.*

Because THESPA is a strike probability only, and does not model the probabilities that a storm will be severe or tornadic, the above results make sense. THESPA is much more reliable in predicting the location of mesocyclones/TVSs (Fig. 26) and with considerable skill.

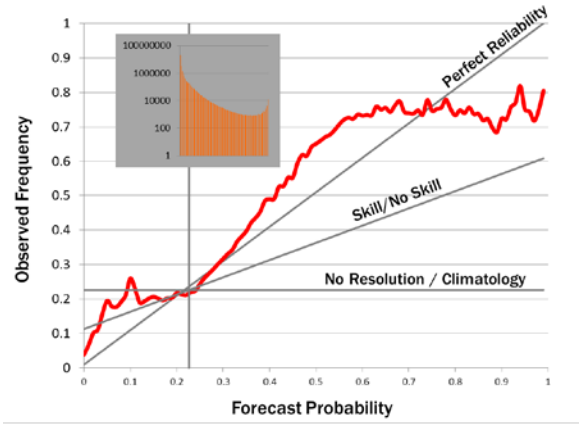


*Figure 26. Reliability diagram for all THESPA probabilities for the mesocyclones/TVSs with all the storms in the afternoon-overnight period. The probability distribution is shown in the inset.*

## 6. DISCUSSION

This new gridded warning verification method addresses the pitfalls of the current NWS warning verification system by consolidating the verification measures into one 2×2 contingency table. The verified hazards can be treated as two-dimensional areas, of which they are – storm hazards do not affect just points or lines. We can also include the correct null forecasts in the measures. This method provides a more robust way to determine location-specific lead times as well as new metrics such as departure time and false alarm time. In addition, the new method will reward spatial and temporal precision in warnings and penalize "casting a wider net" by measuring false alarm areas and false alarm times, which may contribute to a high false alarm perception by the public. The new verification technique also can be used to better address the relative goodness of innovative warning services concepts, such as Threats-In-Motion (TIM) and Probabilistic Hazard Information (PHI).

The new verification system is used to show the benefits of a Threats-In-Motion (TIM) warning concept. Warnings automatically translate downstream based on storm motion until adjusted or cancelled. Because of this, TIM provides equitable lead times for all users downstream of a threat, provides meaningful location-specific information about times of arrival and departure of the threats, and warnings are automatically removed from locations where the threat has passed. TIM warnings would be especially beneficial with GPS enabled mobile warnings applications. The TIM concept is also essential for any future hazard warning system that is based on probabilistic information, because probabilities evolve continuously across time and space. Probabilistic swaths are comprised of higher values near the current threat location, and decreasing values at locations farther downstream from the threat. If probabilistic warnings were only issued at 30-, 45-, or 60- minute intervals, locations farthest downstream of current events would always have lower probabilities. As threats approach these locations, the probabilities should naturally increase. Therefore, TIM must be employed to any warning system that includes a geospatial probabilistic component such as PHI.

Any warning created using a low probability threshold to gain additional lead time will have a higher false alarm ratio. The choice of probability threshold used to warn is dependent on specific users' choice of lead time, false alarm ratio, and their acceptable cost-loss ratio. Users that are more vulnerable to the hazard and/or require greater time to take action to protect themselves or their assets from the hazard may desire the greater lead time, and the trade-off of a higher FAR may be acceptable to their cost-loss model.

Using THESPA as our probabilistic swath model – a strike probability model - we show that probability is more than just a factor of forecaster confidence. There is uncertainty based on the *location* of a current threat. This is modulated by motion uncertainty – for example, will a supercell storm turn to the right and slow down? There is also uncertainty based on the *existence* of a threat. This can be a factor of storm climatology storm environment (e.g., will the storm become tornadic), radar viewing limitations (e.g., is the storm being adequately measured), and of course, forecaster confidence.

## 7. WEB RESOURCES

Additional analysis, figures, animations, and discussion, are available at the following website:

http://tinyurl.com/ewp-thoughts

## 8. ACKNOWLEDGMENTS

This extended abstract was prepared by Gregory J. Stumpf with funding provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

## 9. REFERENCES

Dance,, S., E. Ebert, and D. Scurrah, 2010: Thunderstorm strike probability nowcasting. *J. Atmos. Oceanic Technol.*, **27**, 79–93.

Karstens, C. D., G. Stumpf, C. Ling, L. Hua, D. Kingfield, T. M. Smith, J. Correia, Jr., K. Calhoun, K. Ortega, C. Melick, and L. P. Rothfusz, 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 Hazardous Weather Testbed. *Wea. Forecasting*, **30**, in press.

Lakshmanan, V., T. Smith, K. Hondl, G. Stumpf, A. Witt, 2006: A real-time, three dimensional, rapidly updating, heterogeneous radar merger technique for reflectivity, velocity and derived products. *Wea. Forecasting*, **21**, 802-823.

Marzban, Caren, 1998: Scalar measures of performance in rare-event situations. *Wea. Forecasting*, **13**, 753–763.

NWS, 2011: Service Assessment: The Historic Tornadoes of April 2011, 77 pp.

Stumpf, G. J., T. M. Smith, K. Manross, and D. L. Andra, 2008: The Experimental Warning Program 2008 spring experiment at the NOAA Hazardous Weather Testbed. *Extended Abstracts, 24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., CD-ROM, 8A.1.