A Real-Time Automated Method to Determine Forecast Confidence Associated with Tornado Warnings

John L. Cintineo[1], Travis M. Smith[2,3], Valliappa Lakshmanan[2,3], Kiel Ortega[2,3]

[1]*Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY*
[2]*NOAA/OAR National Severe Storms Laboratory, Norman, OK*
[3]*Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, OK*

Abstract

This paper describes the use of severe weather products derived from the coterminous United States (CONUS) radar network, the lightning detection network, GOES satellites, and model analysis fields to determine the confidence-level of National Weather Service-issued tornado warnings. Severe weather attributes such as low-level shear, reflectivity at -20C and the size of the convective core were extracted (within the geographic and temporal extent of the warning polygons) from the real-time grids produced by the Warning Decision Support System -- Integrated Information (WDSS-II). The initial values of these severe weather parameters at the time the warning was issued were used to determine the conditional probability that a tornado would occur within the spatial and temporal bounds of the warning. Once a warning is issued, it is possible to use this conditional probability to objectively assign a confidence value with the warning in real-time.

1. BACKGROUND

Tornadoes are among the costliest, least predictable atmospheric phenomena, warnings for which had an average lead time of 13 minutes, in 2004 (Erickson, S. A., Brooks, H., 2006). Becoming better at forecasting tornadic storms is not only essential to life and property, but could also decrease the National Weather Service's (NWS) false alarm ratio (FAR), which has economic benefit as well (Erickson, S. A., Brooks, H., 2006). The advent of Doppler radar has provided opportunities to better observe and predict severe weather, via systems such as the National Severe Storms Laboratory's (NSSL) WDSS-II (Lakshmanan, et al 2006). WDSS-II derives products in real-time for the diagnosis of severe weather that aids forecasters in now-casting and warning for severe storms, in particular.

These products are derived from a multi-radar, multi-sensor CONUS merged reflectivity radar network, model and lightning data, and GOES satellite data. The objective of this paper is to investigate archived tornado warnings using these products, and give probabilities to NWS warnings based on different threshold values of the variables.

2. METHODOLOGY

Table – 1 WDSS-II products / storm attributes

Maximum Expected Size of Hail (MESH)	Probability of Severe Hail (POSH)	Severe Hail Index (SHI)	Vertically Integrated Liquid (VIL)
Height of 50dBZ over 253K	Echo top of 18dBZ	Echo top of 30 dBZ	Echo top of 50 dBZ
Area of VIL > 30 kg/m^2	0-2 km Azimuthal Shear	3-6 km Azimuthal Shear	Lowest Level Reflectivity
Reflectivity at 0^{o}C	Reflectivity at -10^{o}C	Reflectivity at -20^{o}C	Maximum Reflectivity
Storm Relative Helicity 0-3km	Storm Relative Flow 9-11km AGL	100mb Avg. CAPE	100mb Avg. CIN
LCL height	IR band-4 temperature	Environmental Shear	Total of 23 products

Twenty-three products (see Table 1) were monitored during each CONUS tornado warning for an unbiased selected 20 days (the same 20 days for each product) from 2 May – 1 July 2008. After this first initial analysis was performed, certain storm attributes were examined in relation to the NWS CONUS tornado warnings from 2 May – 1 July. It should be noted that for 3-6 km Azimuthal Shear (Az. shear), the warnings from 15 May – 2 July were used, since WDSS-II was not processing this attribute before May 15. Also, for 2 May – 10 May, 0-2km Az. shear was replaced with 0-3km Az. shear, since 0-2km Az. shear data was absent from WDSS-II during this time. Environmental data (e.g. CAPE, CIN) is calculated by the 20-km RUC model, which is outputted every hour.

For each archived warning polygon, attribute maximum values inside the warning were calculated for each minute of the warning (minimum values were calculated for IR band-4 temperature, and LCL height). The lifetime maximum (or minimum) for each attribute was also saved after each minute, so by the end of each warning, initial values, lifetime maximums or minimums, and the value at each minute, for each attribute, comprised the dataset. Obviously, if there was no warning valid at a given time, no statistics were produced. Finally, verification of the warnings was performed, using the NWS Storm Prediction Center's storm data, which are still preliminary at the writing of this paper, but suffice for our purposes. Thus, comparisons for verified versus unverified warnings can then be made.

Distributions of initial values and lifetime maximum (or minimum) values were produced for both verified and unverified warnings. The summary of 0-2km Az. shear and vertically integrated liquid (VIL) are presented in this paper. From 2 May – 1 July 2008, there were 1,617 tornado warnings, with an FAR=0.744 and the complement, frequency of hits (FOH) = 0.256. The average warning duration was 38.6 minutes.

2a. 0-2 km Azimuthal Shear

There were 1,237 good warnings (without missing data) for 0-2km Az. shear. Figures 1a) and 1b) show the distributions of initial 0-2km Az. shear (first minute of the warning) for unverified and verified warnings, respectively. Both distributions are skewed to the right, with the unverified warnings skewed more so. This indicates that a large percentage of all tornado warnings had rather weak 0-2km Az. shear. A student's t-test was performed, with unequal sample sizes and unequal variances. The null hypothesis is that the mean initial 0-2km Az. shear for verified warnings is equal to the mean initial 0-2km Az. shear for unverified warnings.

Let a subscript of 1 denote estimators for 0-2km Az. shear for verified warnings, a subscript of 2 denote estimators for 0-2km Az. shear for unverified warnings.

The test statistic is: $t = \dfrac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$

The variance used: $s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

And degrees of freedom: $\mathrm{D.F.} = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$

Using the t-test, we calculated a p-value of 5.23×10^{-11}, which indicates that the initial 0-2km Az. shear means are indeed significantly different for verified and unverified tornado warnings.

Figure 1a)

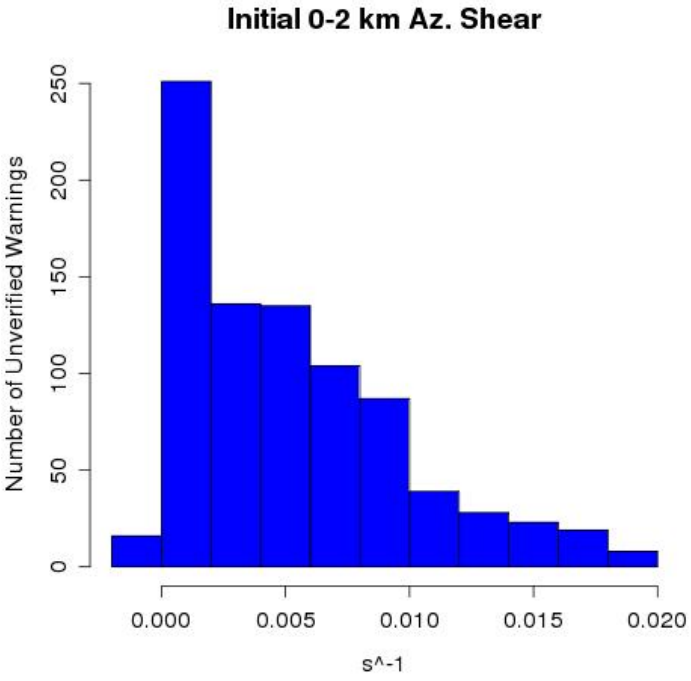

Figure 1b)

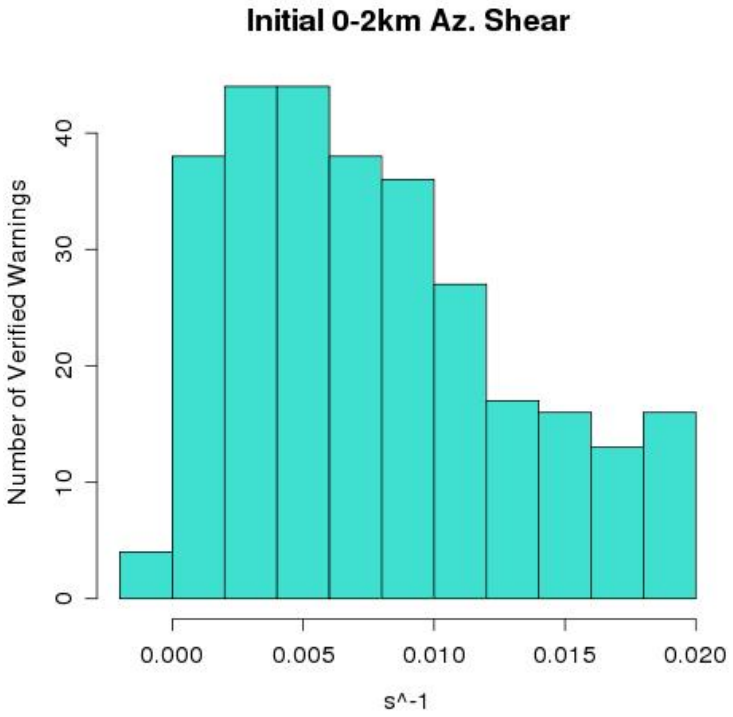

Fig. 1a) The distribution of initial 0-2km Az. shear for *unverified* warnings during the period 2 May -1 July 2008. Mean: 0.0053 s^{-1}, standard deviation: 0.0044 s^{-1}. Fig. 2a) The distribution of initial 0-2km Az. shear for *verified* warnings. Mean: 0.0078 s^{-1}, standard deviation: 0.0053 s^{-1}.

Figure 2a)

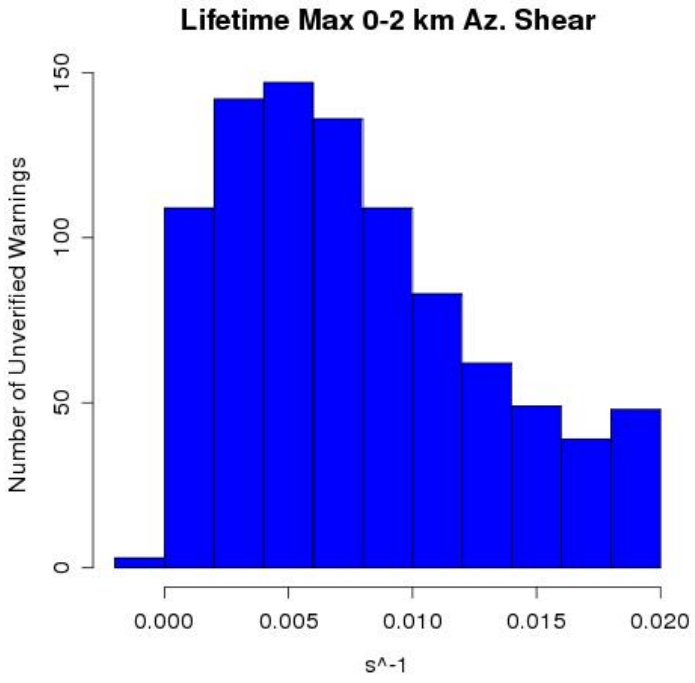

Figure 2b)

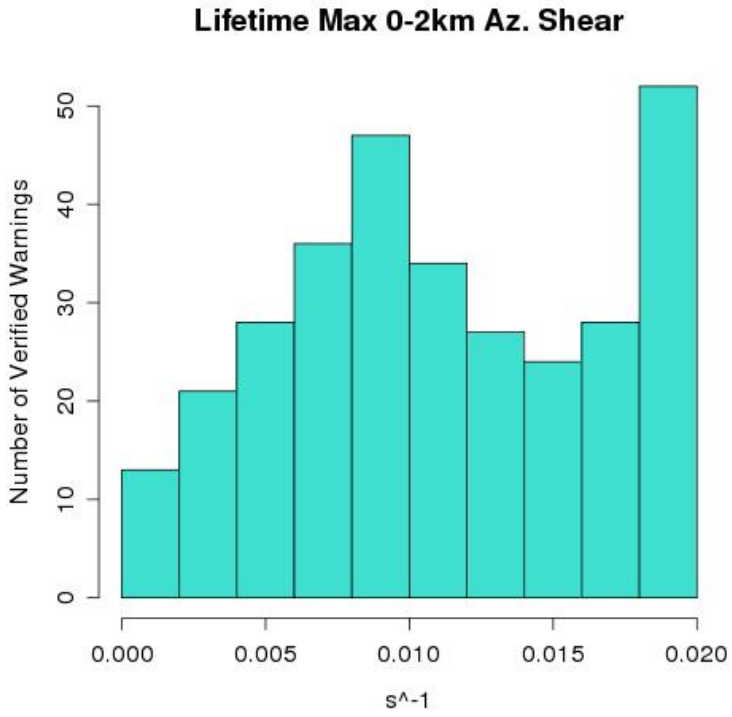

Fig. 2a) The distribution of lifetime maximum 0-2km Az. shear for *unverified* warnings during the period 2 May -1 July 2008. Mean: 0.0078 s^{-1}, standard deviation: 0.0051 s^{-1}. Fig. 2b) The distribution of lifetime maximum 0-2km Az. shear for *verified* warnings. Mean: 0.0109 s^{-1}, standard deviation: 0.0055 s^{-1}.

Figures 2a) and 2b) display the distributions of the lifetime maximum 0-2km Az. shear for unverified and verified warnings, respectively. Both sample means increased, and the distribution for verified warnings appears bimodal, with one mode just under 0.01 s^{-1}, and the other just under 0.02 s^{-1}. This interesting feature could be a result of regional differences in storms (perhaps one mode is for the southeast, and the other for the central Plains), but was not investigated further.

Figure-3 presents a composite time-series of 0-2km Az. shear, for both verified and unverified warnings, over the duration of the warnings. Each bin represents a 5-minute average of the Az. shear, for the 5 minutes prior to but not including the bin value (e.g., bin '5' averages every minute of the warning from initial issuance time to the 5th minute, but not including the 5th minute.). One would expect that initially, the mean 0-2km Az. shear would be about the same for both the verified and unverified warning samples, since one would expect NWS forecasters to warn on approximately the same value for this low-level rotation variable. However, this is not the case. Forecasters are warning on a wide range of storms, with respect to 0-2km Az. shear, as evidenced by the distribution in Figure 1b). While only a fraction of the warnings that verified had weak (say < 0.004 s^{-1}) 0-2km Az. shear, i.e. the probability that a warning had 0-2km Az. shear < 0.004s^{-1}, given it verified is ~10%, forecasters are probably more concerned about their probability of detection (POD) than their FAR. Therefore, they are willing to take a chance and warn on storms with weaker rotation. But logically, the warnings that contain storms with stronger low-level shear tend to verify more, on average.

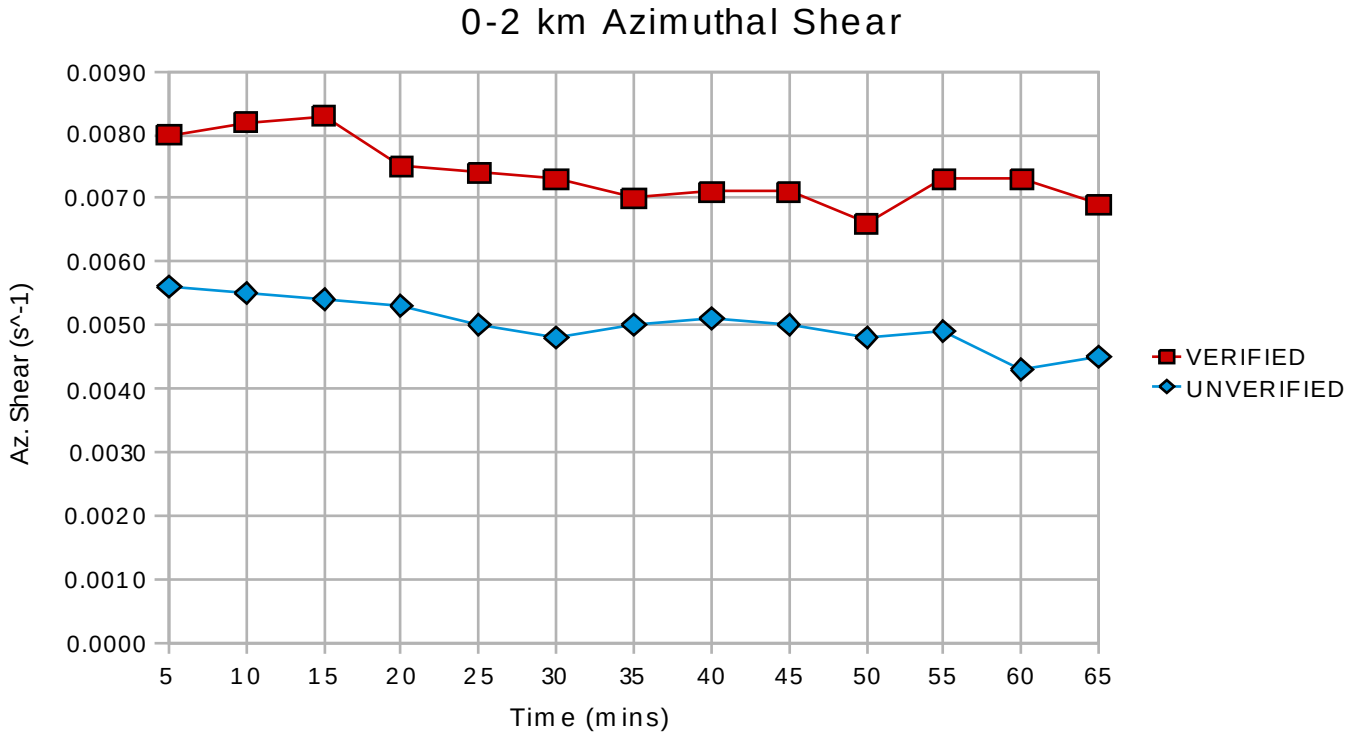

Fig. 3 – A composite time series for 0-2km Az. shear for both verified and unverified warnings. Each time bin represents an average of the five minutes prior to the bin.

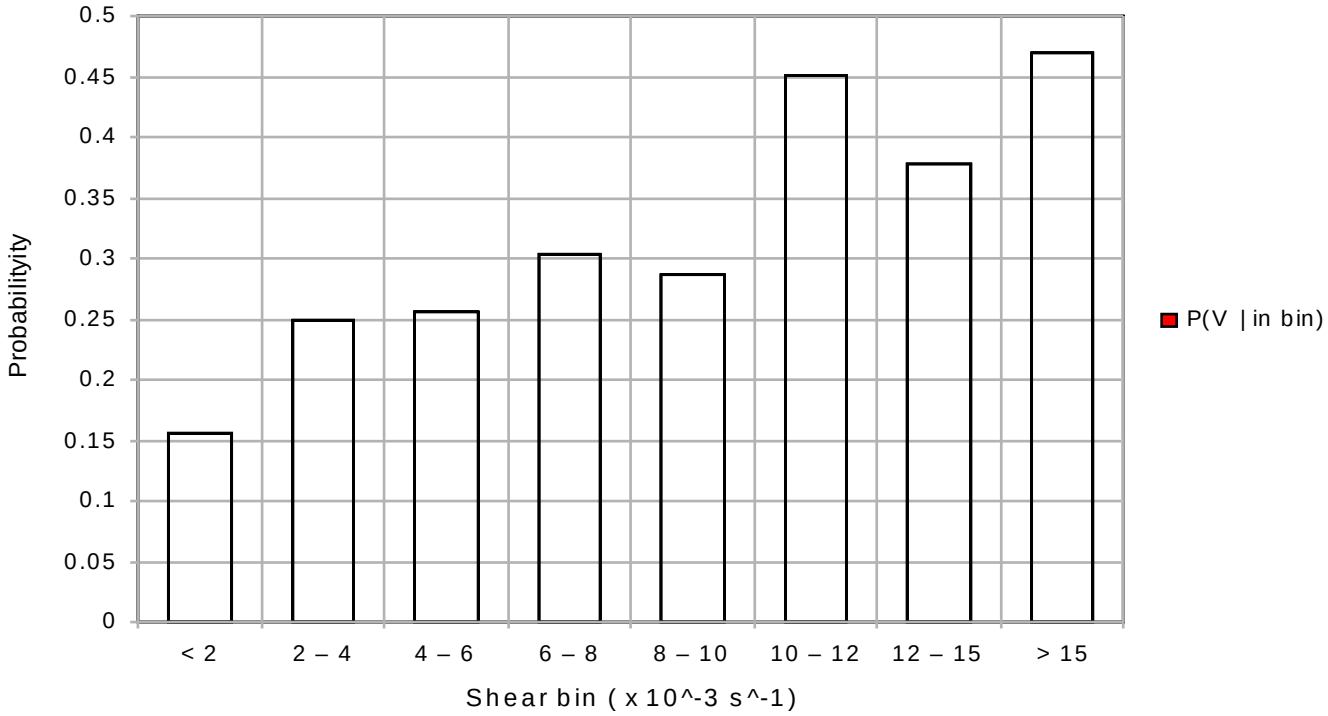

Fig. 4 – The probability that a warning will verify, given an initial 0-2km Az. shear.

Furthermore, from Figure-4, we can see the probability that a warning will verify given a specific range of low-level Az. shear. One can see that the worst probabilities (which happen to correspond to the majority of the warnings issued) are for storms with weak initial shear, and as shear increases, the probability that the warning will verify increases. It should be noted that shear bins are inclusive on the lower bounds, and exclusive on the upper bounds. One other interesting observation is that the probability of verification jumps from around 30% to 45% at the 0.01 s^{-1} threshold, indicating that a value of low-level shear as such may be a key step in issuing a verifying warning. More warning data would help discern the meaning of this jump. If indeed it is a probability jump, a longer climatology of warnings would produce the same feature. If it is simply an artifact of the sample taken, a longer climatology of warnings would most likely smooth out this feature.

2b. Vertically Integrated Liquid

Using multi-radar, multi-sensor merged CONUS reflectivity data, WDSS-II computed the one-minute maximum VIL for each minute of all tornado warnings in the two month period. Again, values were calculated

Figure 5a)

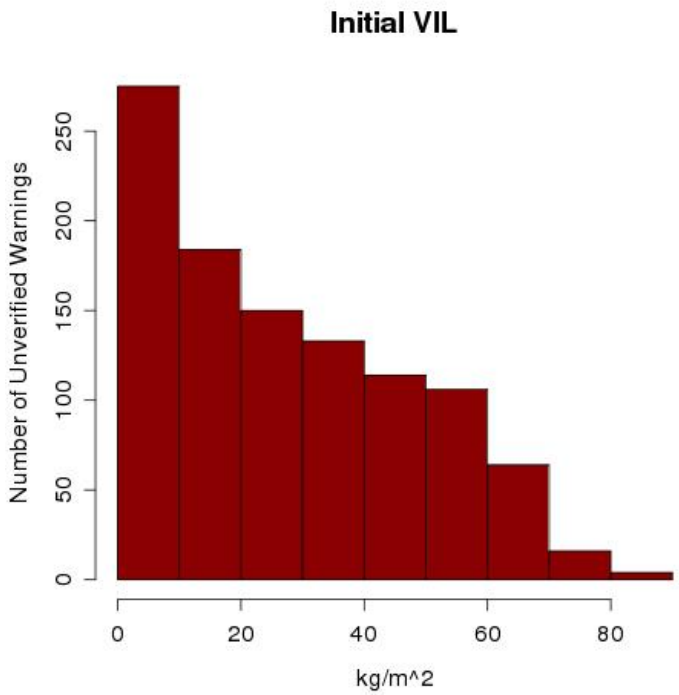

Figure 5b)

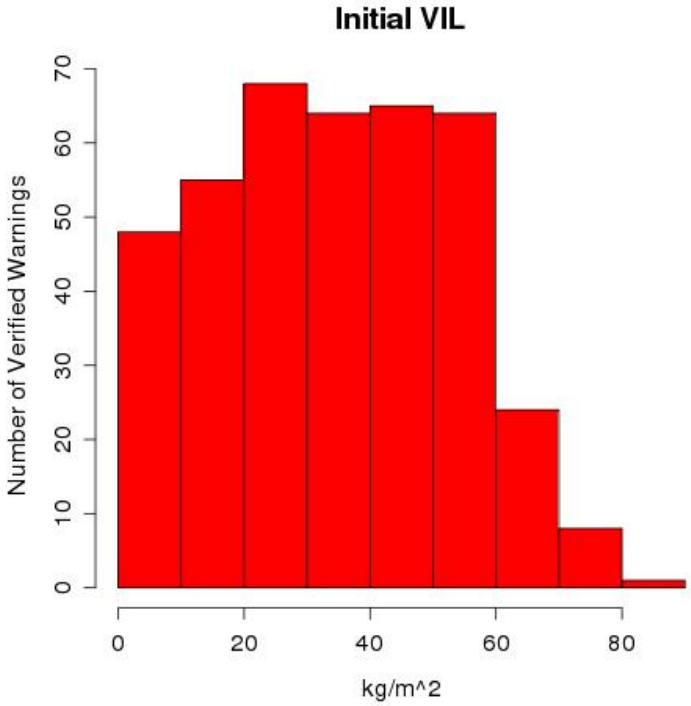

Fig. 5a) The distribution of initial VIL for *unverified* warnings during the period 2 May – 1 July 2008. Mean: 27.76 kg/m^2, standard deviation: 20.46 kg/m^2. Fig. 5b) The distribution of initial VIL for *verified* warnings. Mean: 34.44 kg/m^2, standard deviation: 18.66 kg/m^2.

Figure 6a)

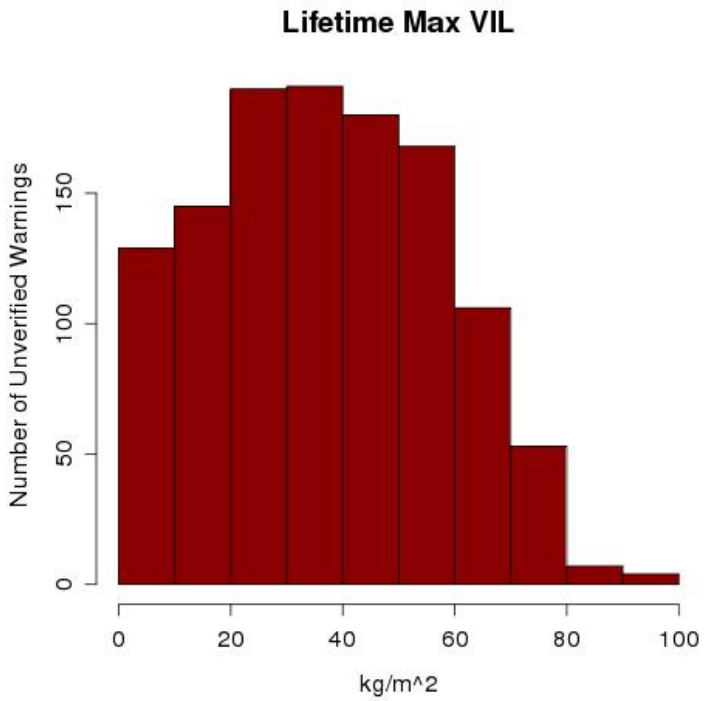

Figure 6b)

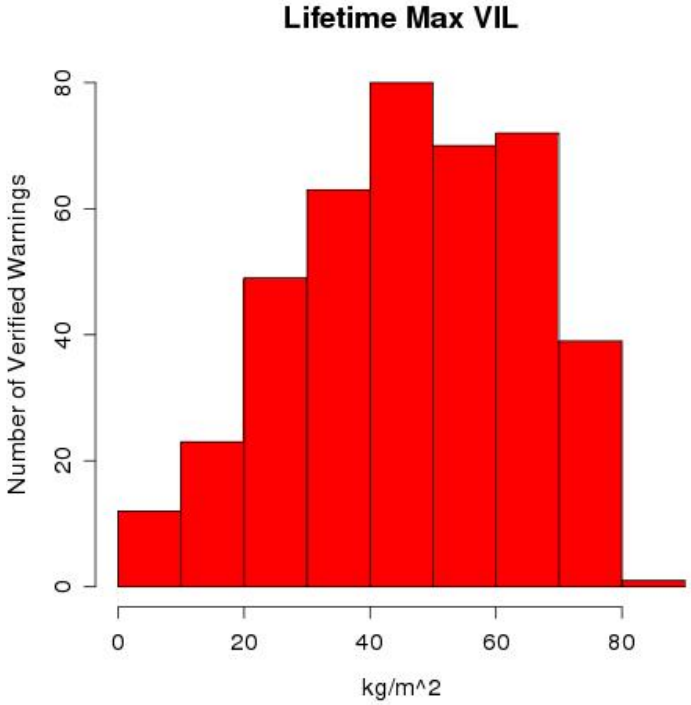

Fig. 6a) The distribution of lifetime maximum VIL for *unverified* warnings during the period 2 May – 1 July 2008. Mean: 37.00 kg/m^2, standard deviation: 20.27 kg/m^2. Fig. 6b) The distribution of lifetime maximum VIL for *verified* warnings. Mean: 46.35 kg/m^2, standard deviation: 18.05 kg/m^2.

entirely inside NWS warning bounds. Figure-5a) displays the distribution of initial VIL for unverified warnings, and figure-5b) shows the initial VIL distribution for the verified warnings. One can see that warnings verified with a wide distribution of initial VIL. Also, a large percentage of warnings were issued with relatively low VIL, as evidenced by the right skewed-ness, especially of the unverified warnings distribution. The mean initial VIL and mean lifetime maximum VIL for the verified warnings are both substantially greater (p-value on the order of 10^{-41}) than those of the unverified warning distributions. Figure-7 presents a full composite time series of VIL, averaged over each five minutes of both verified and unverified warnings.

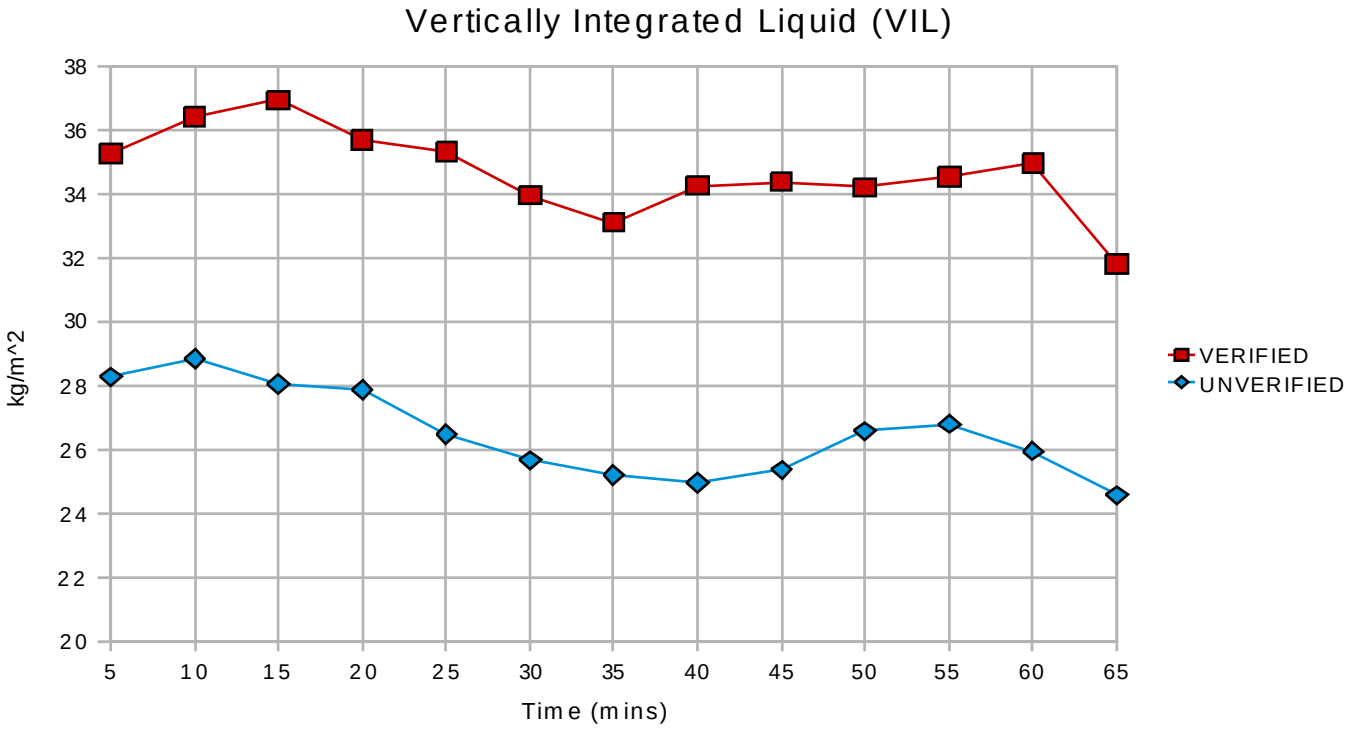

Fig. 7 – A composite time series of Vertically Integrated Liquid (VIL) for both verified and unverified warnings. Each time bin represents an average of the five minutes prior to the bin.

Again, it is rather surprising to see the initial VIL quite different for verified and unverified warnings. Forecasters are warning on a wide range of storms, with respect to VIL, and the storms that tend to verify, have more intense VIL. The reason for such disparity could be that forecasters neglect to use VIL as an indicator for severe storms, or simply, VIL just lacks predictive value for tornadoes. However, the storms that forecasters take a chance on (storms with lower VIL) tend not to verify as much, as evidenced by the consistent lower mean for the unverified warnings time series. Figure-8 quantifies how often a warning is verified, based solely on VIL. One

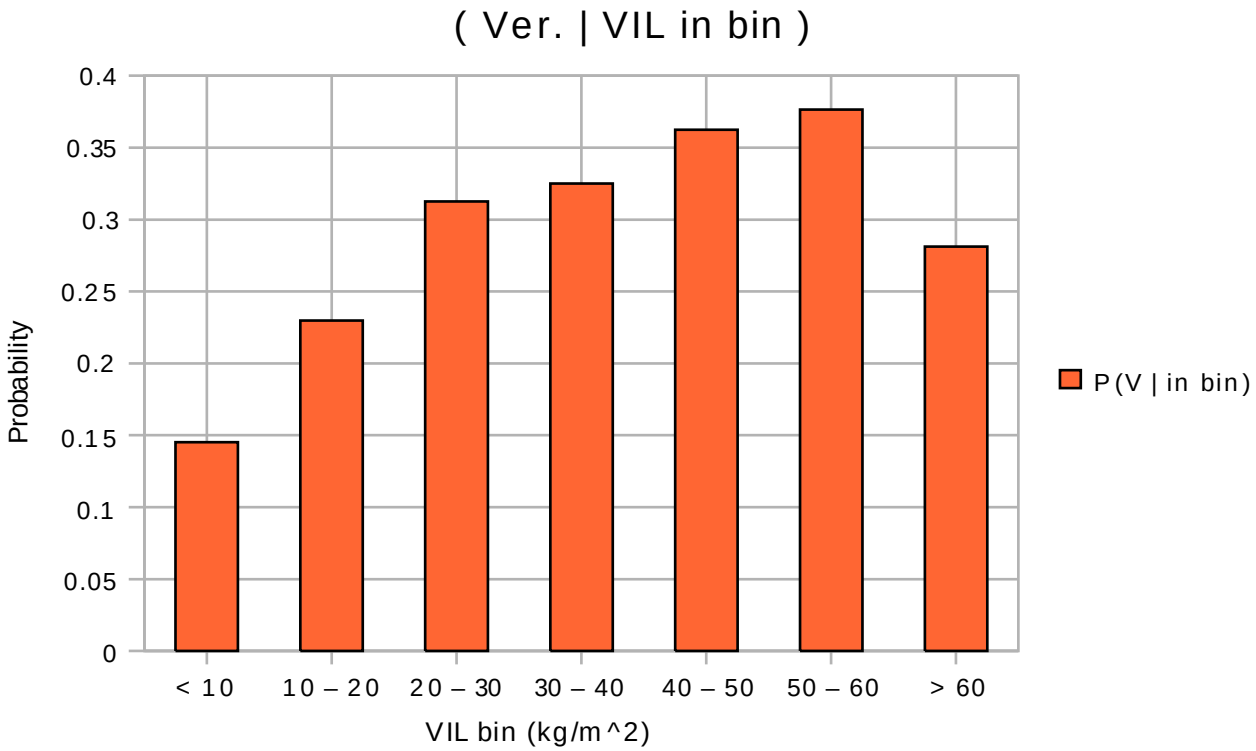

Fig. 8 – The probability that a warning will verify, given an initial VIL.

can see that the probability of verification increases as VIL increases, except for the last VIL bin, with VIL greater than 60 kg/m^2. We attribute this to the lack of storms with a VIL of that magnitude, or again, perhaps this is another hint at VIL's poor predictive capacity. It also is quite apparent that the probabilities of verification for a given VIL are lower across the board than those probabilities of the 0-2km Az. shear, again showing that the low-level Az. shear is a much more effective univariate predictor than VIL.

3. COMPOSITE CONDITIONAL PROBABILITY

It is then possible to put these two products together, to create a conditional probability contingency table, Table – 2. Each cell of the table represents the probability that a tornado will verify, given an initial 0-2km Az. shear and an initial VIL. Both variables were divided into four bins, to ensure a sufficient amount of warnings in order compute a reasonable probability. It is quite noticeable that as low-level Az. shear increases, the probability of verification tends to increase. Conversely, there is no clear pattern with respect to VIL in this bivariate contingency table. As indicated by the distributions of initial 0-2km Az. shear and VIL (Figs. 1a, 1b, 5a, 5b), the cell

Initial 0-2km Azimuthal Shear (s^{-1})

Initial Vertically Integrated Liquid (kg/m^2)	x < 0.004	0.004 <= x < 0.008	0.008 <= x < 0.012	x >= 0.012
y < 20	Ver: 41 Unv: 238 PROB: 0.147	Ver: 10 Unv: 61 PROB: 0.141	Ver: 5 Unv: 34 PROB: 0.128	Ver: 5 Unv: 19 PROB: 0.208
40 > y >= 20	Ver: 29 Unv: 70 PROB: 0.293	Ver: 32 Unv: 49 PROB: 0.395	Ver: 18 Unv: 42 PROB: 0.300	Ver: 24 Unv: 26 PROB: 0.480
60 > y >= 40	Ver: 15 Unv: 41 PROB: 0.268	Ver: 23 Unv: 69 PROB: 0.250	Ver: 26 Unv: 35 PROB: 0.426	Ver: 25 Unv: 27 PROB: 0.481
y >= 60	Ver: 0 Unv: 14 PROB: 0.000	Ver: 9 Unv: 36 PROB: 0.200	Ver: 10 Unv: 15 PROB: 0.400	Ver: 8 Unv: 9 PROB: 0.471

Table – 2: Each cell in this conditional probability contingency table contains the number of verified and unverified tornado warnings, as well as the probability of verification given the bin values of low-level Az. shear and VIL. Only warnings that had non-missing initial low-level Az. shear and initial VIL (after the 1st minute) were used; a total of 1065 warnings.

that encloses both of the lowest bins (top leftmost cell) contains by far the largest number of warnings, as well as one of the worst probabilities of verification. A table as such helps a forecaster quantify his/her own subjective confidence level on issuing a tornado warning. A forecaster could also use a similar product to reduce his/her FAR, if so desired.

4. CONCLUSIONS

The main purpose of this paper was not to isolate the best WDSS-II predictors for tornadoes, but to explain how WDSS-II products can quantify confidence levels for now-casting purposes, namely for the NWS issuing tornado warnings. WDSS-II can automatically assign the probability of verification to a newly issued NWS warning, or even plot that probability, in real-time. A longer climatology of tornado warnings will provide more reliable probabilities of verification, as well as an opportunity to use more than two WDSS-II products as predictors. The larger sample of warnings will ensure that the extreme bins of the multi-dimensional table will be filled with a sufficient number of warnings to calculate a reasonable probability.

Future work would include extending the sample of tornado warnings, and isolating 3 or more of the best WDSS-II products to be used as predictors. It is well known that the environment that a storm initiates in is a crucial factor that forecasters consider when issuing warnings. For example, environments with high instability, and a sufficient lifting mechanism may compel forecasters to issue a warning on the first sight of storm initiation, in which case the reflectivity, VIL, or azimuthal shear at the initial time of the warning would not indicate a strong storm, but rather a storm with the potential to quickly intensify. This could account for the sizeable amount of warnings (both verified and unverified) that were issued by NWS forecasters. Therefore, the environment, with respect to properties such as 0-1km helicity, CAPE, CIN, and LCL height needs to be part of the equation when quantifying forecast confidence. Future study should include a more rigorous investigation of the warning environments. Ideally, several storm environment parameters from the 20-km RUC model would be used to support the WDSS-II radar-derived products when compiling the verification probabilities. Occasionally a tornado warning does not entirely enclose a storm, and consequently WDSS-II would not be processing data on the whole storm. Therefore, another key improvement on this research would be to examine the area just outside of a tornado warning, in order to ensure that WDSS-II products are indeed extracting information from the whole storm, and its environment.

A regional analysis of tornado warnings using WDSS-II would yield tables of confidence levels for different geographical regions of the United States. This type of future study would be especially important, since different regions of the U.S. often have different variable thresholds for severe storms. Another interesting application of this research would be a seasonal comparison of warnings, for the warm and cold season storms, for the same geographical regions. WDSS-II could produce new sets of probabilities for both seasons, and they could then be contrasted to see if there is indeed a difference between warm season storm attributes and cold season storm attributes. Analyzing the difference in tornado warnings that were issued on a storm after an initial warning, and the first warnings on a storm (warnings issued on new convective initiation) could help indicate forecasters' skill at different stages of storm evolution.

A study of especial interest would be to evaluate the differences in storm attributes in severe thunderstorm warnings and tornado warnings, using WDSS-II products. Comparing the forecast confidence levels for both types of warnings could allude to possible threshold values in WDSS-II products, which could in turn help forecasters determine whether to issue a tornado warning or a severe thunderstorm warning.

When now-casting for a severe storm, or multiple threats, forecasters need as much information as quickly as possible to help make the best decision, in terms of issuing a warning or not. Using WDSS-II, a forecaster can get vital storm attribute and environment conditions. If WDSS-II is equipped with the probability of verification based on its products, then the forecaster will also get an objective confidence level, which is a measure of certainty on a particular storm, and another factor in the decision process of issuing a warning.

5. Acknowledgements

The authors would like to thank Owen Shieh, for much thoughtful feedback. This research was supported by an appointment to the National Oceanic and Atmospheric Administration Research Participation Program through a grant award to Oak Ridge Institute for Science and Education.

6. References

Erickson, S. A., and H. Brooks, 2006: Lead time and time under tornado warnings: 1986-2004. *23rd Conference on Severe Local Storms*

Guillot, E., T. M. Smith, V. Lakshmanan, K. L. Elmore, D. W. Burgess, and G. J. Stumpf, 2007: Tornado and Severe Thunderstorm Warning Forecast Skill and its Relationship to Storm Type.

Lakshmanan, V., T. M. Smith, K. Cooper, J. J. Levit, G. J. Stumpf, and D. R. Bright, 2006: High-resolution radar data and products over the Continental United States. *22nd Conference on Interactive Information Processing Systems*, Atlanta, Amer. Meteor. Soc.

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

Lakshmanan, V., T. Smith, G. J. Stumpf, and K. Hondl, 2007: The warning decision support system - integrated information (WDSS-II). Weather and Forecasting 22, 592-608.

Ortega, K. L, and T. M. Smith, 2006: Verification of multi-sensor, multi-radar hail diagnosis techniques. *1st Severe Local Storms Special Symposium, Atlanta, GA, Amer. Meteo. Soc.*

Ortega, K. L., T. M. Smith, G. J. Stumpf, J. Hocker, and L. López, 2005: A comparison of multi-sensor hail diagnosis techniques. *21st Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Amer. Meteo. Soc., P1.11 - CD preprints*.

Witt, A., Eilts, M., G. J. Stumpf, J. T. Johnson, D.E. Mitchell, and K. W. Thomas, 1998: An Enhanced Hail Detection Algorithm for the WSR-88D.