An Enhanced Hail Detection Algorithm for the WSR-88D

ARTHUR WITT, MICHAEL D. EILTS, GREGORY J. STUMPF,* J. T. JOHNSON, E. DEWAYNE MITCHELL,* AND KEVIN W. THOMAS*

NOAA/ERL/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 3 March 1997, in final form 30 January 1998)

ABSTRACT

An enhanced hail detection algorithm (HDA) has been developed for the WSR-88D to replace the original hail algorithm. While the original hail algorithm simply indicated whether or not a detected storm cell was producing hail, the new HDA estimates the probability of hail (any size), probability of severe-size hail (diameter ≥19 mm), and maximum expected hail size for each detected storm cell. A new parameter, called the severe hail index (SHI), was developed as the primary predictor variable for severe-size hail. The SHI is a thermally weighted vertical integration of a storm cell's reflectivity profile. Initial testing on 10 storm days showed that the new HDA performed considerably better at predicting severe hail than the original hail algorithm. Additional testing of the new HDA on 31 storm days showed substantial regional variations in performance, with best results across the southern plains and weaker performance for regions farther east.

1. Introduction

The Weather Surveillance Radar-1988 Doppler (WSR-88D) system contains numerous algorithms that use Doppler radar base data as input to produce meteorological and hydrological analysis products (Crum and Alberty 1993). The radar base data (reflectivity, Doppler velocity, and spectrum width) are collected at an azimuthal increment of 1° and at a range increment of 1 km for reflectivity and 250 m for velocity and spectrum width. Currently, two prespecified precipitation-mode scanning strategies are available for use whenever significant precipitation or severe weather is observed. With volume coverage pattern 11 (VCP-11), the radar completes a volume scan of 14 different elevation angles in 5 min, whereas with VCP-21, a volume scan of 9 elevation angles is completed in 6 min. In either case, the antenna elevation steps from 0.5° to 19.5° (for further details, see Brandes et al. 1991).

In the initial WSR-88D system, one set of algorithms, called the storm series algorithms, was used to identify and track individual thunderstorm cells (Crum and Alberty 1993). The storm series process begins with the storm segments algorithm, which searches along radials of radar data for runs of contiguous range gates having reflectivities greater than or equal to a specified threshold. Those segments whose radial lengths are longer than a specified threshold are saved and passed on to the storm centroids algorithm. This algorithm builds azimuthally adjacent segments into 2D storm components and then builds vertically adjacent 2D components into 3D ''storms.'' The storm tracking algorithm relates all storms found in the current volume scan to storms detected in the previous volume scan. The storm position forecast algorithm calculates a storm's motion vector and predicts the future centroid location of a storm based on a history of the storm's movement. Finally, the storm structure and hail algorithms produce output on the storm's structural characteristics and hail potential.

The initial WSR-88D hail algorithm was developed by Petrocchi (1982). The design is based on identification of the structural characteristics of typical severe hailstorms found in the southern plains (Lemon 1978). The algorithm uses information from the storm centroid and tracking algorithms to test for the presence of seven hail indicators (Smart and Alberty 1985). After testing is completed, a storm is given one of the following four hail labels: positive, probable, negative, or unknown (insufficient data available to make a decision).

Early testing of the hail algorithm showed good performance (Petrocchi 1982; Smart and Alberty 1985). However, subsequent testing by Winston (1988) showed relatively poor performance. Irrespective of its performance, the utility of the hail algorithm is limited by the nature of its output. Since the National Weather Service (NWS) is tasked with providing warnings of severe-size hail (diameter ≥19 mm), it needs an algorithm optimized for this hail size. The aviation community, how-

* *Additional affiliation:* Cooperative Institute for Mesoscale Meteorological Studies, Norman, Oklahoma

Corresponding author address: Arthur Witt, National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069.
E-mail: witt@nssl.noaa.gov

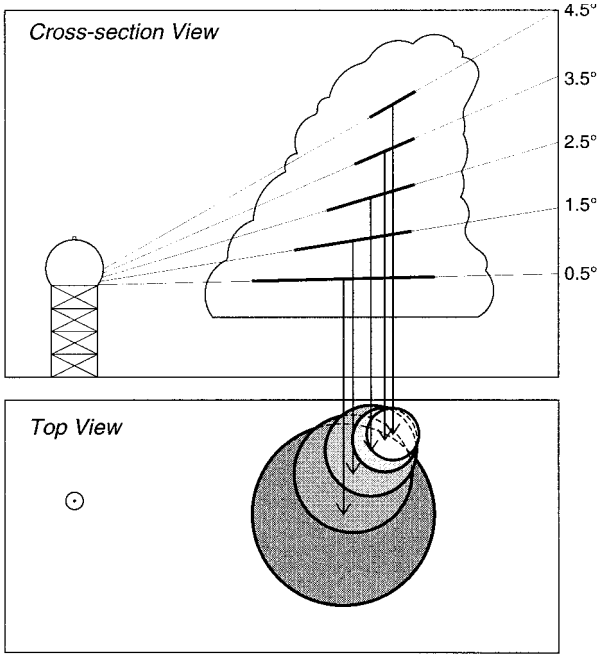

FIG. 1. Diagram illustrating the identification of 2D storm components (thick lines and circles) within a cell by the SCIT algorithm.

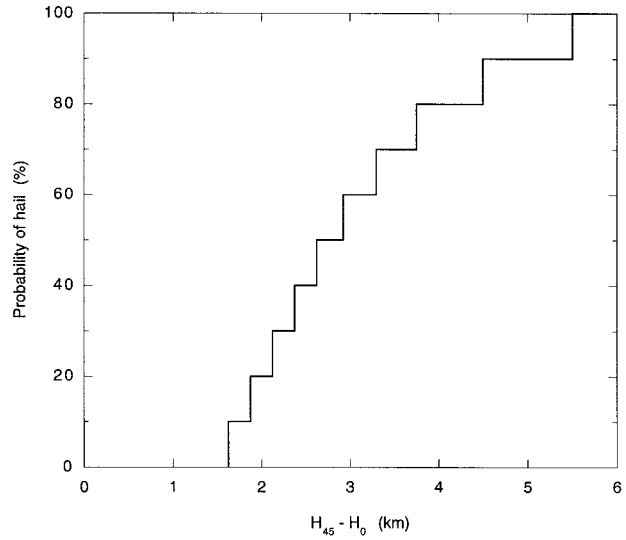

FIG. 2. Probability of hail at the ground as a function of ($H_{45} - H_0$). Here H_{45} is the height of the 45-dB*Z* echo above radar level (ARL), and H_0 is the height of the melting level ARL (derived from Waldvogel et al. 1979).

ever, is interested in hail of any size. Most users would also like an estimate of the maximum expected hail size. Finally, given the general uncertainty involved in discriminating hailstorms from nonhailstorms, or severe hail storms from nonsevere hailstorms, the use of probabilities is advisable.

This has led to the design and development of a new hail detection algorithm (HDA) for the WSR-88D. In place of the previous labels, the new algorithm produces, for each detected storm cell, the following information: probability of hail (any size), probability of severe hail, and maximum expected hail size.

2. Algorithm design and development

The new HDA is a reflectivity-based algorithm and has been designed based upon the demonstrated success of the RADAP II vertically-integrated liquid water (VIL) algorithm (Winston and Ruthi 1986) and techniques used during several hail suppression experiments. The HDA runs in conjunction with the new storm cell identification and tracking (SCIT) algorithm (Johnson et al. 1998). Each cell detected by the SCIT algorithm consists of several 2D storm components, which are the quasi-horizontal cross sections for each elevation angle scanning through the cell (Fig. 1). The height and maximum reflectivity of each storm component are used to create a vertical reflectivity profile for the cell. This information is then used by the HDA to determine a cell's hail potential. To satisfy the different needs of the NWS and the aviation community, the HDA has separate components for detecting hail of any size and severe hail.

a. Detection of hail of any size

To determine the presence of hail of any size, the height of the 45-dB*Z* echo above the environmental melting level is used. This technique has proven to be successful at indicating hail during several different hail suppression experiments (Mather et al. 1976; Foote and Knight 1979; Waldvogel et al. 1979). Using the data presented in Waldvogel et al. (1979), a simple relation between the height of the 45-dB*Z* echo above the melting level and the probability of hail at the ground was derived (Fig. 2).

b. Detection of severe hail

1) SEVERE HAIL INDEX

To determine the presence of severe hail, an approach similar to the VIL algorithm (i.e., vertical integration of reflectivity) was adopted and changes have been made that should improve on its already successful performance. The first change involves moving from a grid-based algorithm to a cell-based algorithm, using output from the SCIT algorithm. The advantage of a cell-based system is that the problem associated with having a hail core cross a grid boundary, and therefore not being accurately measured, is eliminated. The disadvantage is that if an error occurs in the cell identification process, this may cause an error in the HDA.

The second change involves using a reflectivity-to-hail relation, instead of a reflectivity-to-liquid-water relation as VIL does. The reflectivity data are transformed

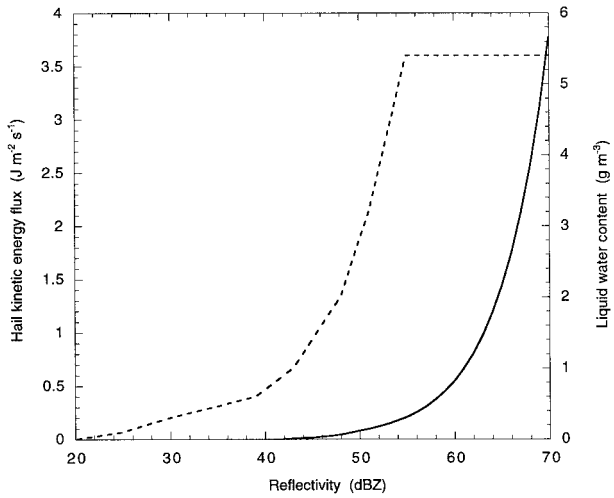

FIG. 3. Plot of hail kinetic energy flux (solid curve), and liquid water content (used to calculate VIL; dashed curve), as a function of reflectivity.

into flux values of hail kinetic energy ($\dot{E}$) (Waldvogel et al. 1978a; Waldvogel et al. 1978b; Federer et al. 1986) by

$$\dot{E} = 5 \times 10^{-6} \times 10^{0.084Z} W(Z), \tag{1}$$

where

$$W(Z) = \begin{cases} 0 & \text{for } Z \le Z_{\mathrm{L}} \\ \dfrac{Z - Z_{\mathrm{L}}}{Z_{\mathrm{U}} - Z_{\mathrm{L}}} & \text{for } Z_{\mathrm{L}} < Z < Z_{\mathrm{U}} \\ 1 & \text{for } Z \ge Z_{\mathrm{U}}. \end{cases}$$

Here Z is in dBZ, $\dot{E}$ in Joules per square meter per second, and the weighting function $W(Z)$ can be used to define a transition zone between rain and hail reflectivities. The default values for this algorithm have initially been set to $Z_{\mathrm{L}} = 40$ dBZ and $Z_{\mathrm{U}} = 50$ dBZ (but are adaptable).[1] From Fig. 3, it can be seen that, whereas the VIL algorithm filters out the high reflectivities associated with hail by having an upper-reflectivity limit of 55 dBZ, the Z–$\dot{E}$ relation functions in the opposite way, using only the higher reflectivities typically associated with hail and filtering out most of the lower reflectivities typically associated with liquid water. Also, $\dot{E}$ is closely related to the damage potential of hail at the ground.

A third change involves using a temperature-weighted vertical integration. Since hail growth only occurs at temperatures <0°C, and most growth for severe hail occurs at temperatures near −20°C or colder (English 1973; Browning 1977; Nelson 1983; Miller et al. 1988), the following temperature-based weighting function is used:

$$W_{\mathrm{T}}(H) = \begin{cases} 0 & \text{for } H \le H_0 \\ \dfrac{H - H_0}{H_{\mathrm{m20}} - H_0} & \text{for } H_0 < H < H_{\mathrm{m20}} \\ 1 & \text{for } H \ge H_{\mathrm{m20}}, \end{cases} \tag{2}$$

where H is the height above radar level (ARL), H_0 is the height ARL of the environmental melting level, and H_{m20} is the height ARL of the −20°C environmental temperature. Both H_0 and H_{m20} can be determined from a nearby sounding or from other sources of upper-air data (e.g., numerical model output).

All of the above leads to the following radar-derived parameter, which is called the severe hail index (SHI). It is defined as

$$\mathrm{SHI} = 0.1 \int_{H_0}^{H_{\mathrm{T}}} W_{\mathrm{T}}(H)\dot{E}\, dH, \tag{3}$$

where H_{T} is the height of the top of the storm cell. In the HDA, SHI is calculated using information from the 2D storm components for the cell being analyzed, with at least two components required for calculation (i.e., SHI values are not calculated for storm cells with just one 2D component). Here $\dot{E}$ is calculated using the maximum reflectivity value for each storm component, and this value is applied across the vertical depth (or thickness) of the storm component. For interior storm components (i.e., those having an adjacent component both above and below them), the vertical depth ΔH_i of the component is given by $\Delta H_i = (H_{i+1} - H_{i-1})/2$. For the top and bottom storm components, $\Delta H_N = H_N - H_{N-1}$ (N being the number of 2D components) and $\Delta H_1 = H_2 - H_1$, respectively. If the height of the base of the storm cell is above H_0, then $\Delta H_1 = (H_1 + H_2)/2 - H_0$. The units of SHI are Joules per meter per second. An example of (3) applied to a storm cell detected by the SCIT algorithm is shown in Fig. 4.

2) INITIAL DEVELOPMENTAL TESTING OF SHI

To determine the utility of SHI as a severe hail predictor, WSR-88D level II data (Crum et al. 1993) were analyzed for 10 storm days from radar sites located in Oklahoma and Florida (Table 1). The process consisted of running the SCIT algorithm and HDA on the radar data and correlating the algorithm output to severe hail reports, with ground-truth verification coming from *Storm Data* (NCDC 1989, 1992). The following analysis procedure was used.

1) A ''hail-truth'' file was created relating hail reports to storm cells observed in the radar data. This involved recording the time and location of the cell that produced the hail report for a series of volume scans before and after the time of the report (e.g., Table 2). Cell locations

[1] These values are lower than those used by Federer et al. (1986), since severe hail is occasionally observed with storms having maximum reflectivities <55 dBZ.

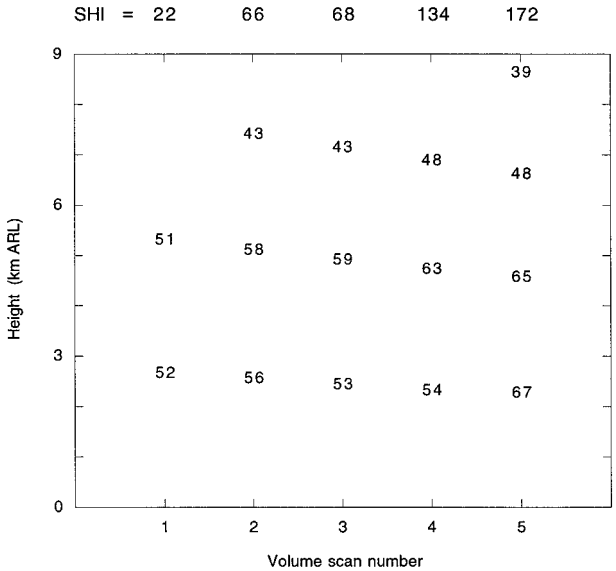

FIG. 4. Sample SHI values (J m^{-1} s^{-1}) for a typical storm cell, along with the corresponding maximum reflectivities (dB*Z*) for each 2D storm component, as identified by the SCIT algorithm for five volume scans. Reflectivity values are plotted at the center height of each component. Here H_0 = 3 km and H_{m20} = 6 km.

were recorded up to 45 min prior to the report time and 15 min after the report time, for those volume scans when the cell had a maximum reflectivity >30 dB*Z*. Storm cells located within the radar's cone of silence (≲30 km) or at ranges >230 km were not analyzed.

2) The algorithm was run using the level II data, and an output file was generated. For each volume scan analyzed, the locations and SHI values of all cells detected by the SCIT algorithm were saved, in decreasing order based on SHI. To avoid the detection of large numbers of relatively small-scale cells within a larger multicellular storm, an adaptable parameter (the minimum separation distance between cell detections) within the SCIT algorithm was set to 30 km. Thus, only the dominant cell within a multicellular storm would be identified by the algorithm.

3) A scoring program was then run using the hail-truth and algorithm output files. The scoring program functions as follows. A ''warning'' threshold is selected. Then, starting with the first volume scan, and continuing until all volume scans are examined, for each storm cell identified, if the SHI value is greater than or equal to the warning threshold, a ''yes'' forecast of severe hail is made for that cell; otherwise, a ''no'' forecast is made. The truth file is scanned to see if the given cell correlates with any of the hail reports. A match occurs if a location entry exists in the truth file for the same volume scan and the distance between the truth-file location and the algorithm location is <30 km. If the cell and a report are related, the entry in the truth file is flagged so that it cannot be associated with any other cells, and the time

TABLE 2. The hail truth file for 2 June 1992. The value on the first line is the number of hail reports for the day. The second (9th) line of values contains the size and time of the first (second) report. The values on lines 3–8 (10–16) are the storm locations (azimuth and range) and volume scan times needed for algorithm scoring.

2	[Hail reports]	
19	2000	[Size (mm), time (UTC)]
Azimuth (°)	Range (km)	Time (UTC)
309	85	1946
310	82	1952
312	78	1958
317	72	2004
317	71	2009
317	71	2015
19	2025	[Size (mm), time (UTC)]
Azimuth (°)	Range (km)	Time (UTC)
328	37	1946
329	39	1952
331	38	1958
330	33	2004
335	32	2009
337	30	2015
337	30	2021

TABLE 1. List of the storm cases analyzed. Here RS is the radar site, BT and ET are the beginning and ending times of data analysis, H_0 is the melting level ARL, NR is the number of hail reports used in the analysis, MS is the maximum reported hail size, NVS is the number of volume scans analyzed, NAP is the total number of algorithm predictions, and MZ is the maximum reflectivity for all the storm cells analyzed. The date corresponds to the beginning time. RS–locations: FDR is Frederick, OK; MLB is Melbourne, FL; OUN is Norman, OK; and TLX is Twin Lakes, OK.

RS	Date	BT (UTC)	ET (UTC)	H_0 (km)	NR	MS (mm)	NVS	NAP	MZ (dbZ)
OUN	1 Sep 1989	1956	0028	4.45	17	51	54	926	69
TLX	11 Feb 1992	2200	0909	2.45	6	25	108	711	64
TLX	17 Feb 1992	0352	0841	2.55	8	25	59	291	57
MLB	25 Mar 1992	2217	0106	3.2	14	76	35	284	75
FDR	19 Apr 1992	0107	0628	3.25	9	102	66	698	69
FDR	28 Apr 1992	1732	0543	3.4	49	70	128	650	72
MLB	28 May 1992	1500	0300	3.8	2	44	132	449	66
MLB	2 Jun 1992	1404	2249	3.85	2	19	91	530	63
MLB	9 Jun 1992	1400	0401	4.3	0	—	128	802	58
MLB	12 Jun 1992	1453	0220	4.2	0	—	132	759	61
Totals	10 days				107		933	6100	

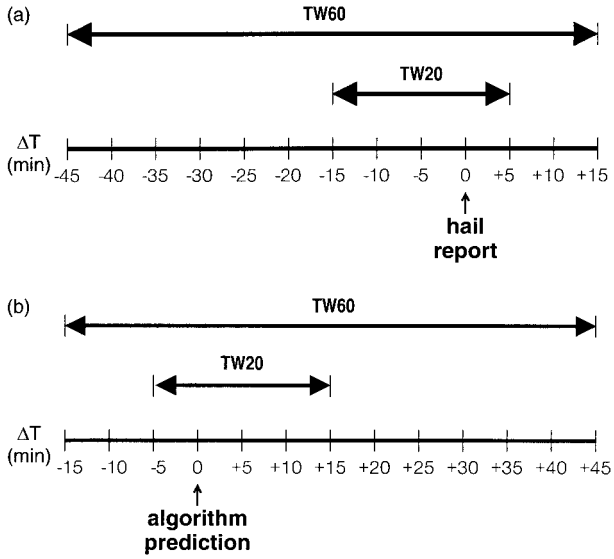

FIG. 5. Diagram of the time window scoring methodology, (a) relative to the time of a hail report and (b) relative to the time of an algorithm prediction.

difference (ΔT) between the report and the current volume scan is calculated (ΔT = volume scan time minus report time). If $T_1 \leq \Delta T \leq T_2$, where T_1 and T_2 define the temporal limits of an analysis ''time window'' (relative to the time of the hail report), then a hit is declared if a yes forecast was made, and a miss is declared if a no forecast was made. If a yes forecast was made, and $\Delta T < T_1$ or $\Delta T > T_2$, or if the prediction is not associated with a hail report, then a false alarm is declared. If an algorithm prediction is associated with more than one hail report, resulting in multiple hits and/or misses (due to overlapping time windows from two or more hail reports), only one hit or miss is counted. Finally, any location entries in the hail-truth file that have not been matched to a storm cell, and fall within the time window of their corresponding report, are counted as misses.

Performance results were determined for two time windows of different lengths. The first time window (TW20) was 20 min in length, with $T_1 = -15$ min and $T_2 = 5$ min (Fig. 5a). The choice of these specific temporal limits was based on the time it takes for large hail (already grown and located at midaltitudes) to fall out of a storm, which is typically up to 10 min (Changnon 1970), and a 5-min buffer zone was added onto both ends of this initial 10-min interval (-10 min $\leq \Delta T \leq$ 0 min) to account for synchronization errors between radar and hail observation times. The second time window (TW60) was 60 min in length, with $T_1 = -45$ min and $T_2 = 15$ min. These temporal limits were chosen based on the time it takes for large hail to both grow and fall out of a storm, which can be up to $\sim$30 min (English 1973), and a 15-min buffer zone was added onto both ends of this initial 30-min interval (-30 min $\leq \Delta T \leq 0$ min) to produce a length similar to that of typical NWS severe weather warnings. An alternate way of visualizing the time windows, relative to the time of an algorithm prediction, is shown in Fig. 5b.

This scoring methodology was used in order to deal with the verification problems caused by the highly sporadic nature of the severe hail reports in *Storm Data*, while still allowing for evaluation of all the algorithm's predictions (Witt et al. 1998). Since many of the reports in *Storm Data* are generated through the verification of NWS severe weather warnings (Hales and Kelly 1985), the reports contained therein will often be on the same time- and space scales as the warnings, which are typically issued for one or more counties for up to 60 min in length. This led to the choice of 60 min as the length of the second time window and was an additional factor in the choice of a large minimum separation distance between cell detections. Thus, for those situations where a storm produces a long, continuous swath of large hail, but hail reports are relatively infrequent (but still frequent enough to verify severe weather warnings), a long time window effectively ''fills in'' the time gap between individual reports. However, since storms can gradually increase in strength before initially producing large hail, and multicellular storms can produce large hail in short, periodic bursts, it would be inappropriate to use just a single, long-time window for algorithm evaluation (because too many misses would be counted in these situations). An additional reason for using a skewed time window (i.e., a larger period before versus after the time of the report) is that this allows for the evaluation (indirectly) of algorithm lead-time capability, which is particularly important to the NWS (Polger et al. 1994).

In the process of building the hail-truth files for this dataset, there were many instances when a hail report either could not be easily correlated to a storm cell based on the radar data (e.g., hail was reported at a specific location and time, with the nearest storm cell 50 km away), or occurred at the edge of a cell, away from the higher reflectivity core ($Z \geq 45$ dBZ). For the 10 storm days analyzed here, there were 115 hail reports in *Storm Data*, of which 33 (29%) did not correlate well with the radar data, if the location and time of the report were assumed to be accurate. One possible solution was to simply discard these reports, but given the general scarcity of ground-truth verification data, this was deemed unacceptable. Instead, an attempt was made to correct such reports. The procedure involved assuming that the location of the report was generally correct (to within a few km), but that the time of the report was in error (up to ± 1 h). The radar data were perused to see if a storm cell had, in fact, passed over the location of the report, within an hour's time of the report and, if so, the original time was changed to correlate best with the radar data and the report was added to the truth file. Of the 33 questionable hail reports, 24 were corrected in this manner. An additional questionable report was added by correcting a typographical error in the location

TABLE 3. Performance results for the new HDA for 17 February 1992 using the 20-min. time window (TW20). Here WT is warning threshold, H is hits, M is misses, FA is false alarms, POD is probability of detection, FAR is false-alarm rate, and CSI is critical success index. POD, FAR, and CSI values are in percent.

WT ($J\ m^{-1}\ s^{-1}$)	H	M	FA	POD (%)	FAR (%)	CSI (%)
10	21	1	57	95	73	27
15	17	5	39	77	70	28
20	16	6	24	73	60	35
25	13	10	12	57	48	37
30	10	13	5	43	33	36
35	6	18	2	25	25	23

$POD = H/(H + M)$ $FAR = FA/(H + FA)$ $CSI = H/(H + M + FA)$

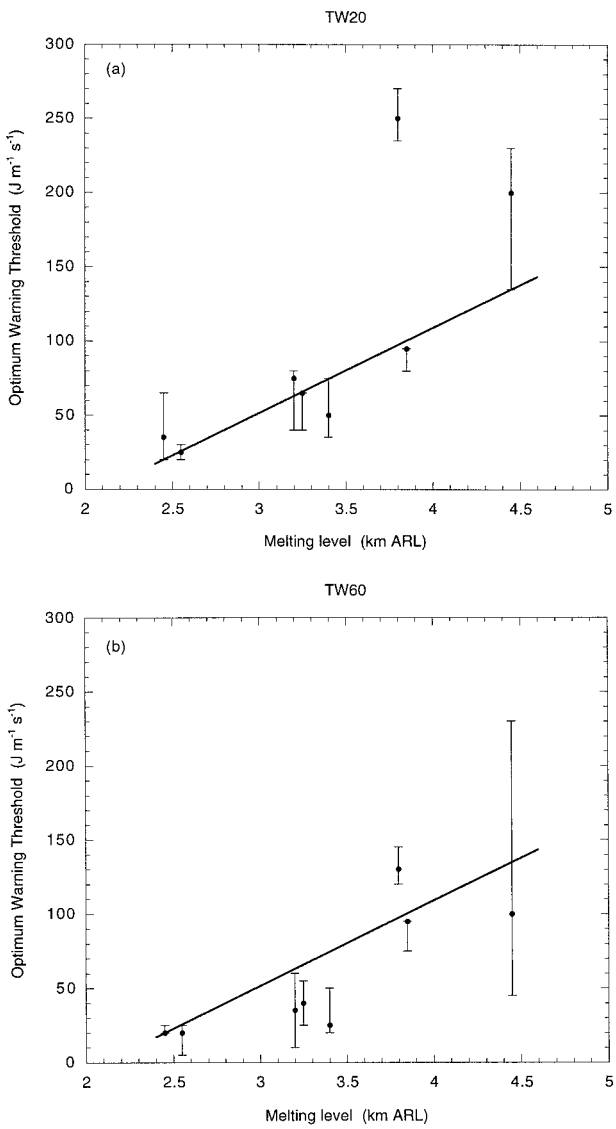

FIG. 6. Optimum warning threshold as a function of the melting level for the 8 days in Table 1 with hail reports for (a) TW20 and (b) TW60. Solid circles correspond to the highest CSI. Vertical bars represent the range of warning thresholds with a CSI within 5 percentage points of the maximum value (i.e., nearly optimal). The sloping line is the warning threshold selection model.

of the report. Eight of the original 115 reports were ultimately discarded.

Using the time-window scoring method mentioned above, performance results were generated for all the storm days analyzed using a multitude of different warning thresholds. As an example, the results for the 17 February 1992 case (for TW20) are shown in Table 3. Similar tables of scoring results (not shown) were generated for each of the other storm days, and the warning threshold producing the highest critical success index (CSI) was noted. For the 8 storm days when severe hail was observed, it was found that the optimum warning threshold (leading to the highest CSI) was highly correlated with the melting level on that day [linear correlation coefficients of 0.78 and 0.81 for TW20 and TW60, respectively (Fig. 6)]. From these results, a simple warning threshold selection model (WTSM) was created and is defined as

$$WT = 57.5H_0 - 121, \tag{4}$$

where WT ($J\ m^{-1}\ s^{-1}$) is the warning threshold and H_0 (km) is measured ARL.[2] If WT $< 20\ J\ m^{-1}\ s^{-1}$, then WT is set to $20\ J\ m^{-1}\ s^{-1}$.

Using (4), a new set of performance results were generated (Table 4). For each severe hail day, the number of hits is lower, and false alarms higher, for TW20 versus TW60. Misses are higher by a factor of at least 2 for TW60 versus TW20, except for the 28 May 1992 case. This all leads to higher probability of detection (POD) and false-alarm rate (FAR) values for TW20 versus TW60, except for the 28 May 1992 case, where POD values are identical. The corresponding CSI values are higher on 3 days and lower on 5 days for TW20 versus TW60. This raises the question of which set of results is more likely representative of actual algorithm performance. For POD, the values from TW20 are probably more accurate, as some of the no forecasts that are being

[2] It should be noted that, at the present state of algorithm development, a flat earth is assumed. This will undoubtedly lead to errors in the WT calculations for those radar sites where terrain height varies substantially within 230 km. Hopefully, future enhancements to the algorithm will include site-specific terrain models for those locations where this correction is needed.

TABLE 4. Performance results for the new HDA. Cases are listed by increasing warning threshold (melting level). For those days with severe hail reports, the first row of values are for TW20, with the second row for TW60. (See Table 3 for definition of terms.)

Date	WT (J m^{-1} s^{-1})	*H*	*M*	FA	POD (%)	FAR (%)	CSI (%)
11 February 1992	20	16	1	33	94	67	32
		24	7	25	77	51	43
17 February 1992	26	13	10	11	57	46	38
		18	32	6	36	25	32
25 March 1992	63	30	9	18	77	38	53
		40	29	8	58	17	52
19 April 1992	66	16	12	21	57	59	31
		22	28	15	44	41	34
28 April 1992	74	94	39	32	71	25	57
		118	98	8	55	6	53
28 May 1992	97	5	0	10	100	67	33
		10	0	5	100	33	67
2 June 1992	100	3	3	6	50	67	25
		5	8	4	38	44	29
12 June 1992	120	0	0	5	—	100	0
9 June 1992	126	0	0	0	—	—	—
1 September 1989	134	40	20	71	67	64	31
		71	44	40	62	36	46
Overall		217	94	207	70	49	42
		308	246	116	56	27	46

TABLE 5. Same as Table 4, but for the original WSR-88D hail algorithm using a warning threshold of ''probable'' (i.e., both ''probable'' and ''positive'' indications are used to make positive hail forecasts).

Date	*H*	*M*	FA	POD (%)	FAR (%)	CSI (%)
11 February 1992	0	19	0	0	—	0
	0	36	0	0	—	0
17 February 1992	6	18	4	25	40	21
	6	45	4	12	40	11
25 March 1992	25	14	24	64	49	40
	35	44	14	44	29	38
19 April 1992	24	3	78	89	76	23
	35	12	67	74	66	31
28 April 1992	103	26	43	80	29	60
	131	71	15	65	10	60
28 May 1992	5	0	37	100	88	12
	10	0	32	100	76	24
2 June 1992	3	3	28	50	90	9
	6	7	25	46	81	16
12 June 1992	0	0	81	—	100	0
9 June 1992	0	0	21	—	100	0
1 September 1989	53	7	204	88	79	20
	101	10	156	91	61	38
Overall	219	90	520	71	70	26
	324	225	415	59	56	34

counted as misses with TW60 are likely occurring at times when storms are not producing large hail. Conversely, for FAR, the values from TW60 are probably more accurate, since the yes forecasts that are counted as false alarms with TW20, but not with TW60, do correspond to a known severe hail event, the full extent of which is unknown due to deficiencies in the verification data (Witt et al. 1998). Consequently, the CSI values (for either time window) are likely understating actual algorithm performance.

For comparison, performance results were also generated for the original WSR-88D hail algorithm (using the same procedure given above) and are shown in Table 5. Although the overall number of hits and misses (and POD values) are nearly the same for the two algorithms, the original WSR-88D hail algorithm produces many more false alarms, with a FAR much higher than that of the new algorithm. Comparing CSI values for the days with severe hail, and the number of false alarms for the days with no severe hail reports, the new algorithm outperforms the original algorithm on 9 of the 10 days.

3) DEVELOPMENT OF A PROBABILITY FUNCTION

Given the general success of SHI and the WTSM at predicting severe hail (overall CSI values >40%), the final stage of development was to implement an appropriate probability function. Since the dataset used for development thus far was quite small, it was decided that the initial probability function should be fairly simple in nature to avoid overfitting the data. Candidate functions were first developed (by trial and error) using test results (for TW60) from only 2 storm days, those with the lowest and highest melting levels, and their calibration (for all 10 storm days) was determined using reliability diagrams (Wilks 1995). This rather limited initial analysis led to a surprisingly good (for this developmental dataset) probability function, which is given by

$$\mathrm{POSH} = 29 \ln\left(\frac{\mathrm{SHI}}{\mathrm{WT}}\right) + 50, \qquad (5)$$

where POSH is the probability of severe hail (%), POSH values <0 are set to 0, and POSH values >100 are set to 100. Despite the continuous nature of (5), actual algorithm output probabilities are rounded off to the nearest 10%, in order to avoid conveying an unrealistic degree of precision. Note that when SHI = WT, POSH = 50%. The reliability diagram of (5) applied to all 10 storm days is shown in Fig. 7.

c. *Prediction of maximum expected hail size*

The SHI is also used to provide estimates of the maximum expected hail size (MEHS). Using data from the 8 severe hail days shown in Table 1, along with data from Twin Lakes on 18 June 1992, (yielding a total of 147 severe hail reports), an initial model relating SHI to maximum hail size was developed. The process involved comparing SHI values with observed hail sizes. For each hail report in the dataset, the maximum value of SHI within TW20 was determined. A scatterplot of these SHI values versus observed hail size is shown in Fig. 8. One thing that is clearly seen in Fig. 8 is the common practice of reporting hail size using familiar

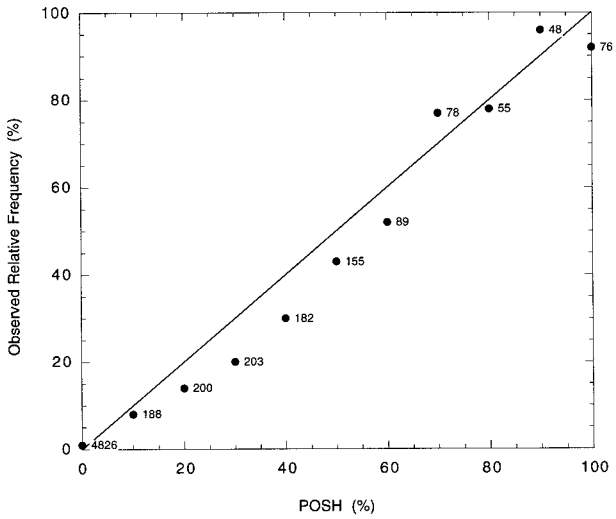

FIG. 7. Reliability diagram for the probability of severe hail for the 10 storm days in Table 1. Numerals adjacent to the plotted points indicate the number of forecasts for that POSH value. The diagonal line represents perfect reliability.

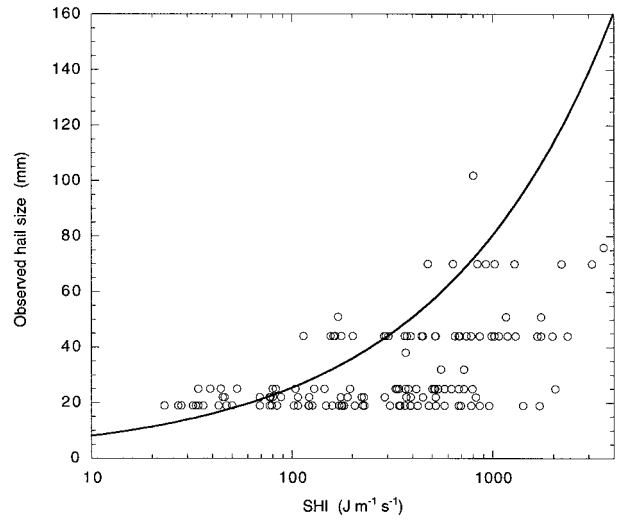

FIG. 8. Scatterplot of SHI vs observed hail size for 147 hail reports from 9 storm days. The plotted curve is the MEHS prediction model.

circular or spherical objects (e.g., various coins or balls) as reports tend to be clustered along discrete sizes. Concerning the relationship between SHI and hail size, it is apparent that the minimum and average SHI (for the different common size values) increase as hail size increases. However, there does not appear to be an upper-limit cutoff value for SHI as hail size increases. This is likely due to the fact that a storm producing very large hail will almost always be producing smaller diameter hail at the same time (often falling over a larger spatial area than the very large hail), and this smaller (but still severe-sized) hail will also usually be observed and reported.

Since the hail-size model being developed is meant to forecast *maximum* expected hail size, it was developed such that around 75% of the hail observations would be less than the corresponding predictions. As was the case with development of the probability function for POSH, it was decided that the initial hail-size prediction model should also be fairly simple in nature. This led to the following relation:

$$\mathrm{MEHS} = 2.54(\mathrm{SHI})^{0.5}, \qquad (6)$$

with MEHS in millimeters. Equation (6) is also shown in Fig. 8. Comparing (6) with the hail-size observations shows that it meets the 75% goal mentioned above and is close to 75% for each of the three distinct size clusters (Table 6). Again, to avoid conveying an unrealistic degree of precision, actual algorithm output size values are rounded off to the nearest 6.35 mm (0.25 in.).

3. Performance evaluation

a. Hail of any size

Evaluating the performance of the probability of hail (POH) parameter was difficult due to the lack of available ground-truth verification data (for hail of any size). However, during the summer months of 1992 and 1993, the National Center for Atmospheric Research (NCAR) conducted a hail project in the high plains of northeastern Colorado to collect an adequate dataset for algorithm verification (Kessinger and Brandes 1995). As part of the hail project, both the new HDA and the original hail algorithm were run using reflectivity data from the Mile High Radar (Pratte et al. 1991), a prototype Next Generation Weather Radar (NEXRAD) located 15 km northeast of Denver. Given the highly detailed nature of the special verification dataset that was collected, it was possible to score algorithm performance on an individual volume scan basis, instead of the time-window method that was developed for use with *Storm Data.* Performance results are summarized in Kessinger et al. (1995), with detailed results given in Kessinger and Brandes (1995). Two pertinent overall results are repeated here. Using 50% as a warning threshold for the POH parameter, and verifying against hail observations of any size, the following accuracy measures were obtained: POD = 92%, FAR = 4%, and CSI = 88%. A similar evaluation of the original hail algorithm using ''probable'' as a warning threshold gave

TABLE 6. Hail-size observations compared to model predicted sizes for 9 storm days.

Hail size (mm)	Number of observations	Percentage of observations less than model (%)	Average SHI ($\mathrm{J\ m^{-1}\ s^{-1}}$)
19–33	99	77	325
33–60	37	70	724
>60	11	73	1465
All	147	75	511

TABLE 7. List of the additional storm cases analyzed. See Table 1 for definition of terms. New locations (RS): DDC is Dodge City, KS; LSX is St. Louis, MO; LWX is Sterling, VA; MKX is Milwaukee, WI; MPX is Minneapolis, MN; NQA is Memphis, TN; and OKX is New York City, NY.

RS	Date	BT (UTC)	ET (UTC)	H_0 (km)	NR	MS (mm)	NVS	NAP	MZ (dbZ)
TLX	4 March 1992	2018	0549	2.5	7	44	100	529	62
MLB	6 March 1992	1438	0305	3.7	6	44	137	537	71
OUN	8 March 1992	1500	0637	3.05	78	89	155	1322	69
MLB	7 June 1992	1232	0955	4.2	0	—	221	1012	62
MLB	8 June 1992	1319	1131	4.3	0	—	220	958	59
TLX	18 June 1992	1831	0407	4.1	45	70	94	643	70
TLX	19 June 1992	1706	0051	4.2	12	44	81	577	71
LSX	10 August 1992	1748	0159	4.25	0	—	86	830	64
MLB	11 August 1992	1956	0444	4.3	0	—	99	571	62
MLB	20 August 1992	2038	0646	4.3	1	19	120	735	65
LSX	26 August 1992	1917	1524	4.0	0	—	175	2092	65
MLB	29 August 1992	1254	0558	4.15	0	—	167	935	58
MLB	1 September 1992	1221	0451	4.05	3	25	183	840	64
OUN	20 September 1992	2049	0802	3.95	3	38	111	1045	67
LWX	16 April 1993	1223	0943	2.65	10	44	224	1046	62
DDC	5 May 1993	1950	0547	3.45	25	70	94	430	63
DDC	2 June 1993	2150	0842	3.55	52	152	105	772	70
LSX	8 June 1993	1949	1629	3.75	2	19	170	1163	67
LSX	13 June 1993	1918	0945	3.85	0	—	149	160	68
LSX	19 June 1993	1758	0516	3.95	0	—	155	2237	66
LSX	30 June 1993	1651	0131	4.25	12	102	68	176	71
MLB	9 July 1993	1248	0403	4.15	6	38	151	693	64
MLB	10 July 1993	1249	0641	4.12	9	38	190	987	65
MLB	9 August 1993	1231	0815	4.2	7	25	198	734	65
LSX	14 April 1994	2246	1915	3.45	15	51	193	941	67
NQA	26 April 1994	1751	0400	3.7	11	44	119	341	68
NQA	27 April 1994	1623	0449	3.7	30	44	152	2248	70
OKX	20 June 1995	1823	0600	4.2	18	70	110	107	70
MKX	15 July 1995	1352	0357	4.2	5	76	158	624	63
MPX	9 August 1995	0052	0902	4.6	5	64	80	336	67
MKX	9 August 1995	0930	2234	4.55	2	44	141	537	64
Totals	31 days				364		4406	26 158	

these results: POD = 74%, FAR = 5%, and CSI = 72%.

b. Severe hail

To provide an independent test of SHI, the initial WTSM, and the initial probability function used to calculate the POSH parameter, additional testing was done using the same analysis procedures presented in section 2b. Since this algorithm testing occurred in phases spanning a period of several years, case selection was largely determined by the availability of WSR-88D level II data at the time of testing. Despite these constraints, it was still possible to obtain radar data from numerous different sites across the United States (Table 7).

To test the accuracy of SHI and the WTSM, performance statistics were again generated using the WTSM to produce categorical forecasts of severe hail for each day listed in Table 7, with results shown in Tables 8 and 9. Table 8 gives performance statistics for each individual day, and Table 9 shows overall performance statistics for cases grouped together into different geographical regions.

Algorithm performance varied widely from one storm day to another. For null cases (i.e., days with no severe hail reports), the best result possible was zero false alarms (e.g., 8 June 1992). However, on some days (e.g., 10 August 1992) the HDA produced many false alarms. For those days with reported severe hail, the HDA had CSI values that varied from a low of 3% (on 8 June 1993) to a high of 78% (on 20 June 1995). Except for two days, the HDA had POD values ≥50% (for TW20). Conversely, FAR values (for TW60) varied greatly, from a low of 0% (on 20 June 1995) to a high of 96% (on 8 June 1993). Of particular interest are the two days from Memphis (26 and 27 April 1994). The large-scale synoptic pattern was nearly identical for these two days, but algorithm performance was quite different. On 26 April 1994, the HDA performed quite well, producing a CSI of 62% (for TW60). However, on 27 April 1994 (the most active storm day in the dataset, in terms of the number of algorithm predictions), the algorithm produced a very large number of false alarms, resulting in markedly poorer performance. The reasons for this large difference in performance are not known. Comparison of overall test results between the independent and developmental datasets (Table 8 vs Table 4) shows an

TABLE 8. Same as Table 4, but for 31 additional storm days.

Date	WT ($J\ m^{-1}\ s^{-1}$)	*H*	*M*	FA	POD (%)	FAR (%)	CSI (%)
4 March 1992	23	11	7	53	61	83	15
		30	19	34	61	53	36
16 April 1993	31	11	20	7	35	39	29
		17	58	1	23	6	22
8 March 1992	54	159	42	140	79	47	47
		242	117	57	67	19	58
5 May 1993	77	65	9	97	88	60	38
		116	35	46	77	28	59
14 April 1994	77	23	6	42	79	65	32
		33	29	32	53	49	35
2 June 1993	83	101	18	57	85	36	57
		143	51	15	74	9	68
6 March 1992	92	18	3	16	86	47	49
		21	8	13	72	38	50
26 April 1994	92	25	3	57	89	70	29
		56	9	26	86	32	62
27 April 1994	92	99	25	505	80	84	16
		227	71	377	76	62	34
8 June 1993	95	3	3	108	50	97	3
		4	9	107	31	96	3
13 June 1993	100	0	0	21	—	100	0
20 September 1992	106	7	4	56	64	89	10
		14	12	49	54	78	19
19 June 1993	106	0	0	4	—	100	0
26 August 1992	109	0	0	23	—	100	0
1 September 1992	112	8	2	33	80	80	19
		13	6	28	68	68	28
18 June 1992	115	99	13	73	88	42	54
		139	47	33	75	19	63
10 July 1993	116	21	15	81	58	79	18
		38	42	64	48	63	26
29 August 1992	118	0	0	0	—	—	—
9 July 1993	118	13	8	56	62	81	17
		29	31	40	48	58	29
7 June 1992	120	0	0	13	—	100	0
19 June 1992	120	27	9	84	75	76	23
		50	24	61	68	55	37
9 August 1993	120	19	5	25	79	57	39
		34	15	10	69	23	58
20 June 1995	120	28	8	0	78	0	78
		28	25	0	53	0	53
15 July 1995	120	6	8	14	43	70	21
		8	17	12	32	60	22
10 August 1992	123	0	0	69	—	100	0
30 June 1993	123	26	7	36	79	58	38
		57	20	5	74	8	70
8 June 1992	126	0	0	0	—	—	—
11 August 1992	126	0	0	2	—	100	0
20 August 1992	126	3	1	20	75	87	13
		6	6	17	50	74	21
9 August 1995a	141	6	4	39	60	87	12
		12	7	33	63	73	23
9 August 1995b	144	11	1	18	92	62	37
		22	7	7	76	24	61
Overall		789	221	1749	78	69	29
		1339	665	1199	67	47	42

increase in both the POD and FAR and a decrease in the CSI.

On a regional basis, substantial performance variations exist (Table 9). The POD values exhibit the smallest amount of regional variation, with large differences in regional FAR values. These FAR differences are the primary factor leading to the corresponding regional variations in CSI (i.e., lower relative FAR corresponds with higher relative CSI). Also shown in Table 9 are overall POD values for two larger hail-size thresholds. Except for Florida (FL), the POD increases as hail size increases.

Another test of the accuracy of the WTSM is to deter-

TABLE 9. Regional performance results. For each region, the first row of overall POD, FAR, and CSI values are for TW20, with the second row for TW60. Here NR is the number of hail reports, NAP is the number of algorithm predictions, SP is southern plains (DDC, OUN, TLX), FL is Florida (MLB), MR is Mississippi River (LSX, NQA), and NUS is northern United States. (MKX, MPX, OKX). The last two columns are for larger hail diameters (D). POD, FAR, and CSI values are in percent.

Region	Number of days	NR	NAP	Overall POD (%)	Overall FAR (%)	Overall CSI (%)	POD D >25 mm (%)	POD D >51 mm (%)
All	31	364	26 158	78	69	29	87	96
				67	47	42		
SP	7	222	5318	82	54	41	92	99
				71	29	55		
FL	10	32	8042	71	75	23	71	—
				57	57	32		
MR	9	70	10 188	80	83	16	89	100
				73	64	32		
NUS	4	30	1604	71	58	36	79	84
				56	43	39		

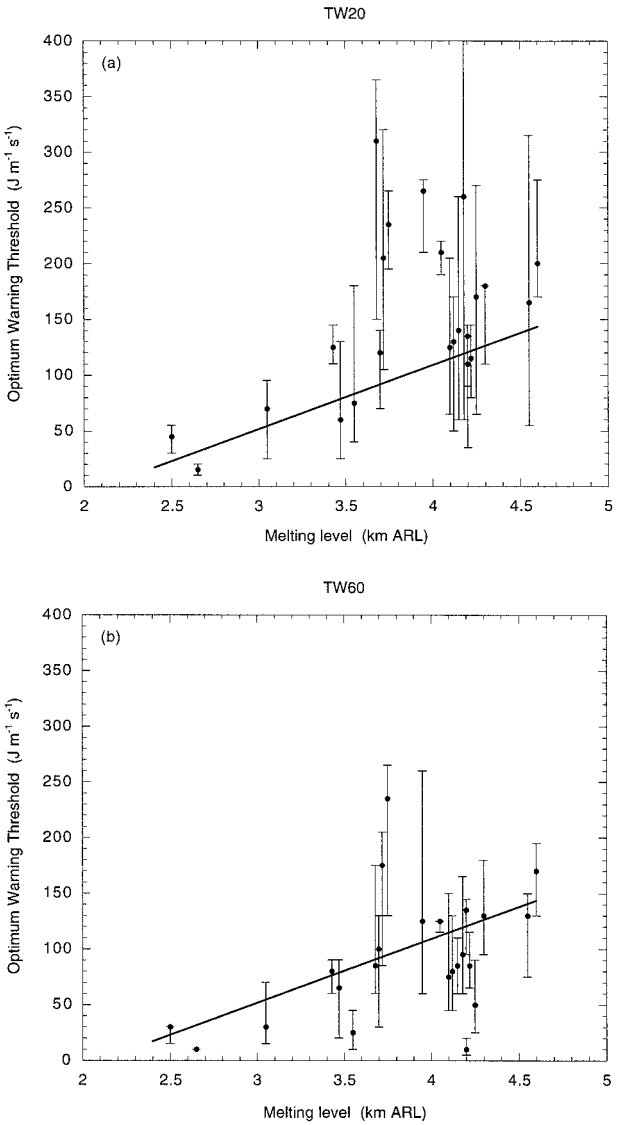

FIG. 9. Same as Fig. 6, except for the 31 storm days in Table 8.

mine if it remains highly correlated with the melting level and, if so, whether the initial model equation is still the best one to use. Therefore, for each severe hail day in Table 7, the optimum warning threshold was calculated and plotted versus the day's melting level (Fig. 9). For both time windows, most of the days (74% for TW20 and 78% for TW60) have optimum warning thresholds (including the five-point range bars) on or close to the WTSM.[3] For those days with optimum warning thresholds not close to the WTSM, these were all higher than the WTSM for TW20 and, except for one day, were all lower than the WTSM for TW60. Regional variations are shown in Figs. 10 and 11. For all regions except the Mississippi River (MR), there is a generally good match between the WTSM and the observed optimum warning thresholds.

To evaluate the POSH parameter, reliability diagrams were again used. Figure 12 shows the reliability diagram for all the days listed in Table 7. Although Fig. 7 showed a slight overforecasting bias (for medium-range probabilities) for the developmental dataset, Fig. 12 shows a pronounced overforecasting bias for the independent dataset. However, this overforecasting bias varies dramatically for the different regions (Fig. 13). For the southern plains, there is little bias and very good calibration. For the northern United States, there is a considerable overforecasting bias for probabilities of 20%–60%, and also 80%, with the remaining probability values showing good calibration. However, for FL, and especially the MR region, large overforecasting biases exist. For these two regional datasets, the initial probability function developed for the POSH parameter shows very poor calibration, and suggests the need for regionally dependent definitions of the POSH parameter.

The effect of population density on algorithm performance was investigated in a very limited study involving the two Wisconsin cases. In addition to the anal-

[3] Close to the WTSM is defined as being within 20 J m^{-1} s^{-1}.

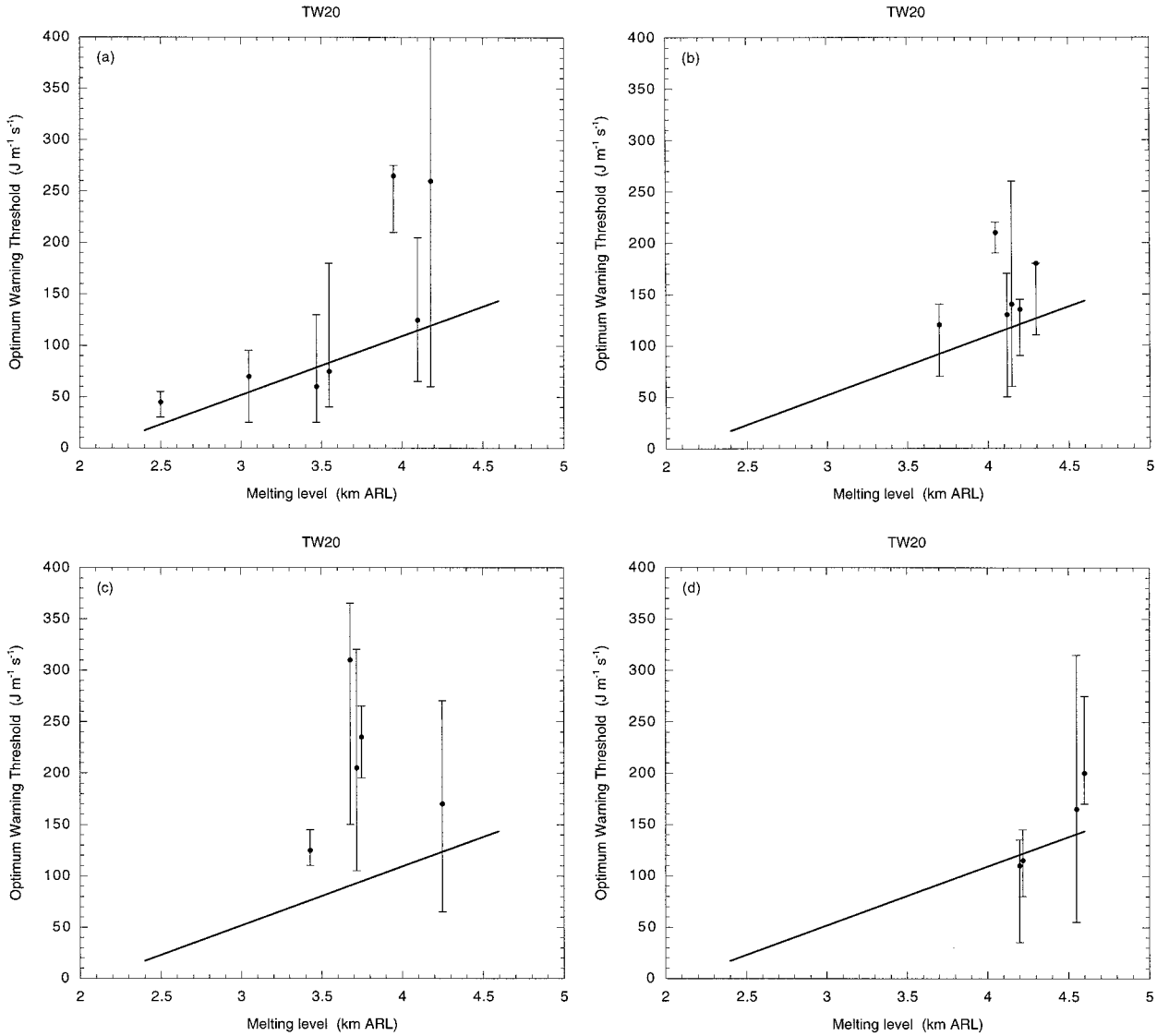

FIG. 10. Same as Fig. 9, except for TW20 subdivided into four different regions: (a) southern plains, (b) Florida, (c) Mississippi River, and (d) northern United States.

ysis results presented in Table 8, a second evaluation, limited to storms occurring over the Milwaukee (MKE) metropolitan area, was done (Table 10). As would be expected, limiting the analysis domain to only the MKE area greatly reduced the number of algorithm predictions available for evaluation. However, given that the remaining storm events occurred over an urban area (high population density), it is much less likely that a severe weather event would go unreported, compared to the full domain. Thus, any false alarms produced by the algorithm are more likely to be valid, and not simply because the storm occurred over an area with few, if any, storm spotters. Comparing the full and MKE domain performance results does, in fact, show large differences in the FAR values, with superior CSI values for the MKE domain. And the CSI (and POD) can be increased to even higher values for the MKE domain by lowering the algorithm's warning threshold by 33% (MKE2). What these results seem to indicate is that some, and possibly many, of the false alarms shown in Tables 3–5 and 8 (and also affecting Figs. 6, 7, and 9–13) may be fictitious. Thus, some of the large overforecasting bias seen in Figs. 12 and 13 could be due to underreporting of actual severe hail events. However, because of the small amount of data analyzed here, further investigation is needed in order to validate this hypothesis. Initial results from a larger study of this issue (Wyatt and Witt 1997) also show improved algorithm performance for higher population density areas.

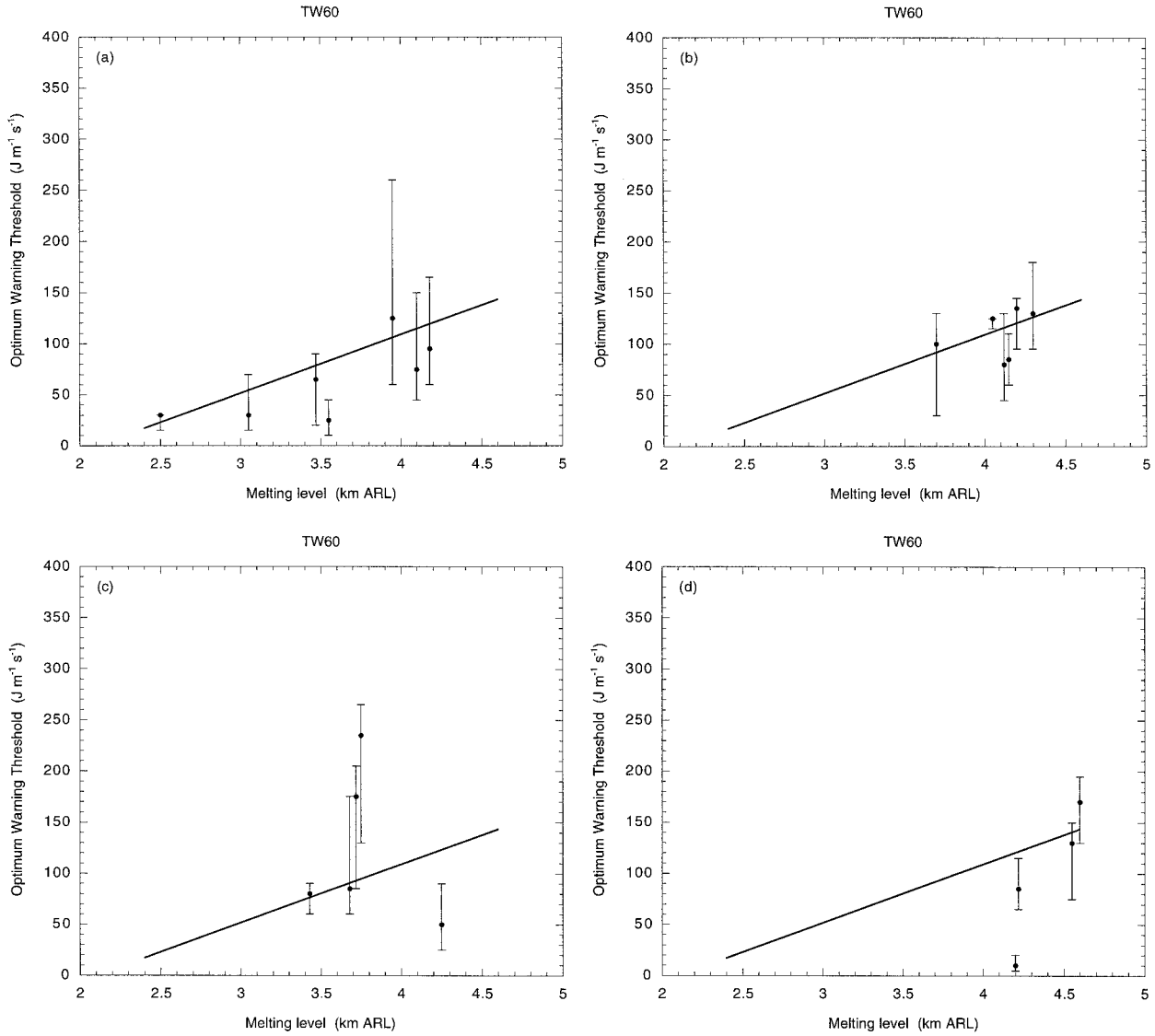

FIG. 11. Same as Fig. 10, except for TW60.

c. *Maximum hail size*

An independent evaluation of SHI as a hail-size predictor was done using hail reports from the days given in Table 7 (minus the reports from 18 June 1992, which were used in the initial development of the hail-size model), along with some supplemental reports (diameters >4 cm) from the days shown in Table 11 (yielding a total of 314 reports).[4] Once again, the maximum value of SHI within TW20 was determined and plotted versus the size of the hail report (Fig. 14). Comparing Fig. 14 to Fig. 8, it is apparent, for sizes >33 mm, that the average value of SHI has decreased substantially, resulting in a smaller percentage of observed sizes greater than the MEHS model curve (Table 12). Also evident is an increased vertical stacking of the observations, thus reducing the discrimination capability of SHI as a hail-size predictor.

4. Discussion

The new WSR-88D HDA attempts to do considerably more than its original counterpart. Instead of simply providing a single, categorical statement on whether or not a storm is producing hail, it tries to determine the potential hail threat from multiple perspectives and provide quantitative guidance to end users. However, the

[4] Some of the hail reports listed in Table 7 were not usable for the size evaluation, because they occurred during time periods without complete radar data, or at ranges >230 km or ≤30 km. They were usable in the other evaluations because of the time-window scoring methodology.

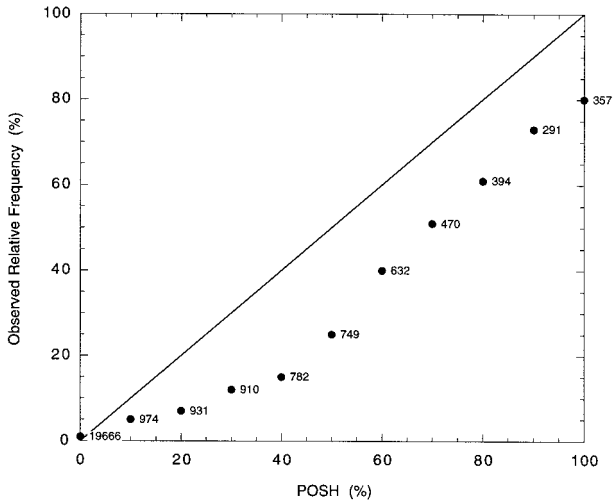

FIG. 12. Same as Fig. 7, except for the 31 storm days in Table 8.

TABLE 10. Same as Table 4, but for the two MKX cases for different analysis domains (AD). NAP is the number of algorithm predictions and MKE2 refers to the case where the warning threshold has been reduced by 33%.

AD	NAP	*H*	*M*	FA	POD (%)	FAR (%)	CSI (%)
Full	1161	12	12	53	50	82	16
		20	24	45	45	69	22
MKE	37	2	2	0	50	0	50
		2	5	0	29	0	29
MKE2	37	4	0	1	100	20	80
		5	2	1	71	17	63

ability to properly design and develop an algorithm like the new HDA (i.e., one that is empirical in nature and produces detailed quantitative information) depends greatly on the quality and quantity of ground-truth data available for development and testing. Inadequacies and errors in the ground-truth database will have a corre-

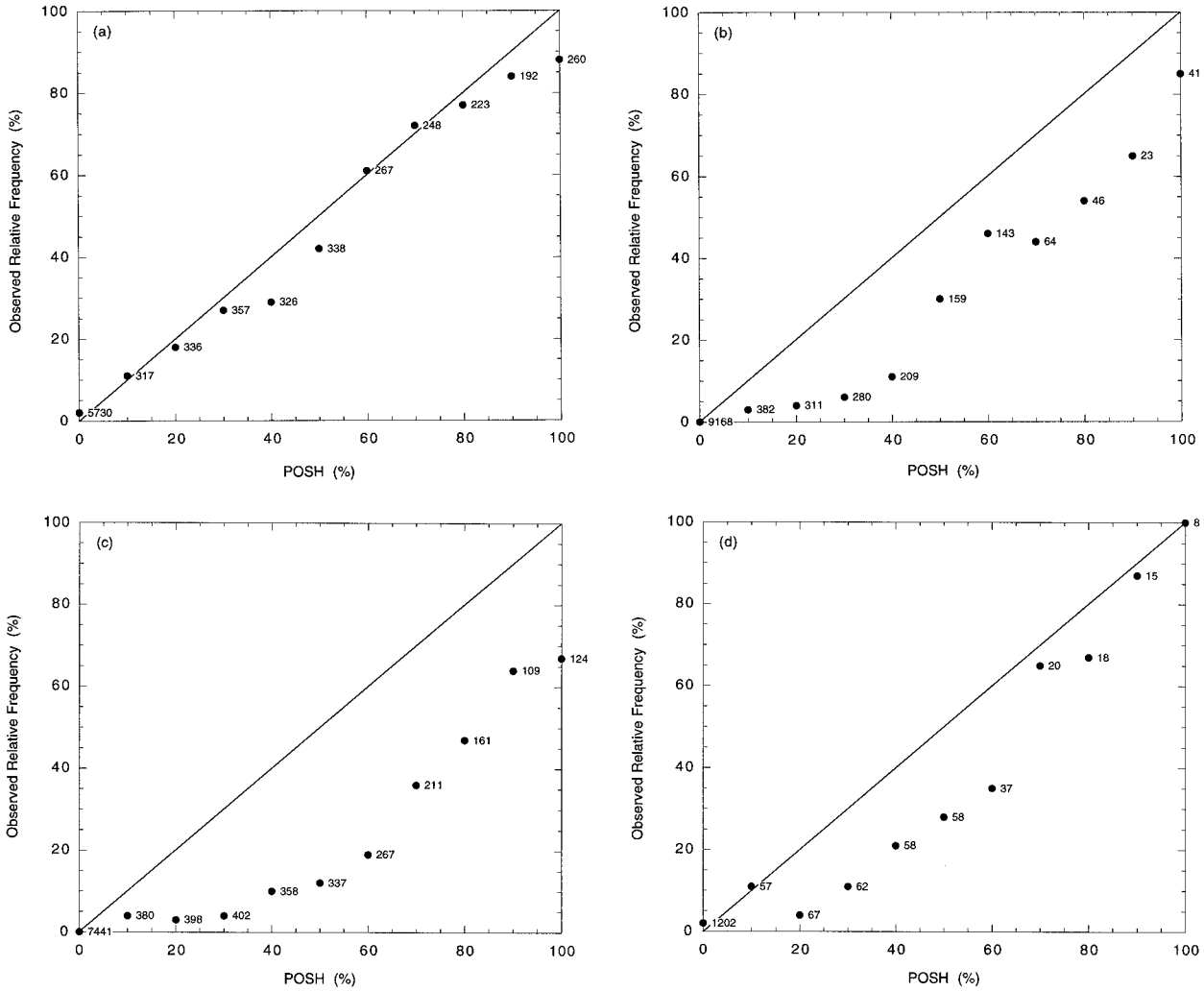

FIG. 13. Same as Fig. 12, except subdivided into four different regions: (a) southern plains, (b) Florida, (c) Mississippi River, and (d) northern United States.

TABLE 11. List of the additional storm cases analyzed to increase the number of very large hail reports. See Table 1 for definition of terms. Additional RS–location: IWA is Phoenix, AZ.

RS	Date	BT (UTC)	ET (UTC)	H_0 (km)	NR	MS (mm)
OUN	12 April 1992	0049	0559	3.25	2	70
OUN	19 April 1992	0039	0103	3.25	2	89
OUN	11 May 1992	2013	2252	3.45	6	89
FDR	14 May 1992	0229	0249	3.6	1	70
OUN	2 September 1992	2333	0317	3.7	4	102
FDR	29 March 1993	2032	0512	3.12	6	89
FDR	2 May 1993	0033	0259	3.2	2	89
IWA	24 August 1993	2204	2216	4.82	1	44
Totals	8 days				24	102

sponding negative impact on algorithm design and performance. For development and testing of the new HDA, verification data has come both from special field projects (such as NCAR's hail project) and from *Storm Data*. Field project datasets are limited in scope but are generally of high quality. On the other hand, *Storm Data* provides severe weather information for the entire United States, but the information is less detailed and often less accurate.

The probability of hail parameter was both developed and tested using special field project data. Hence, ground-truth deficiencies and errors should be minimal. The test results from Kessinger et al. (1995) show that the POH parameter performs very well in Colorado. However, it should be noted that the development and testing of the POH parameter exclusively involved data collected in a ''high-plains''-type of geographical environment. Therefore, it is possible, and perhaps even likely, that the performance of the POH parameter will be poorer in other regions of the United States.

Unlike the POH parameter, the probability of severe hail parameter was developed using *Storm Data* for ground-truth verification. One thing that is obvious from the results presented in section 3 is that the POSH parameter performs considerably better in the southern plains than in other parts of the United States. There are several reasons why this may be so. One potential reason has to do with ground-truth verification efficiency, that is, the percentage of actual severe weather events that are observed and reported to the NWS. From the performance statistics shown in Table 9, it is clear that regional variations in CSI are largely a function of variations in FAR. The cause of this variation in the FAR is unknown. However, the information that appears in *Storm Data* is largely the result of NWS severe weather warning verification efforts (Hales and Kelly 1985) and thus is a function of both severe weather climatology and verification efficiency. Now, if verification efficiency was constant across the United States, then the regional differences in algorithm performance could be attributed solely to differences in severe weather climatology. But verification efficiency is not constant across the United States (Hales and Kelly 1985; Crowther and Halmstad 1994). Some NWS offices put a greater emphasis on severe weather verification than do others, and population density varies dramatically across the United States. Therefore, regional differences in algorithm performance are a function of both differences in severe weather climatology and differences in verification efficiency, with the largest impact of verification efficiency being on the FAR statistic.

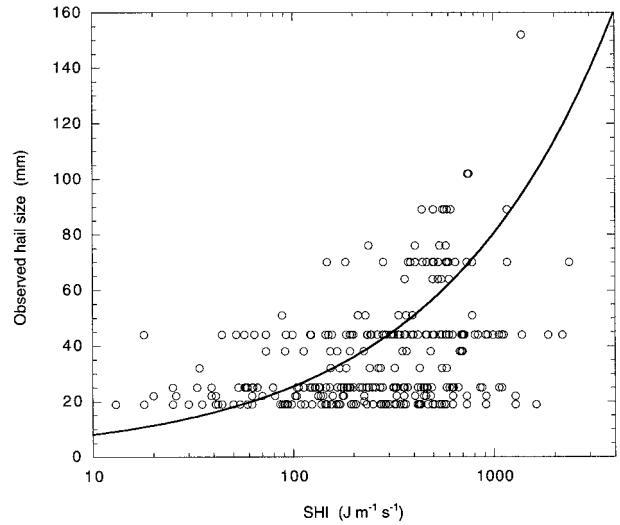

FIG. 14. Same as Fig. 8, except for 314 hail reports from 30 storm days.

As it is, both of these factors are likely affecting the regional performance statistics. Considering the severe weather climatology aspect, large hail is simply more common in the Great Plains compared to other parts of the United States (Kelly et al. 1985). There, hailstorms are often fairly long-lived and produce longer hailswaths (Changnon 1970), making observation of a single hailfall event more likely. Considering the verification efficiency aspect, since NWS offices in the southern plains tend to have extensive severe weather spotter networks, warning verification efforts often lead to many hail reports during severe storm events (Table 7). Changnon (1968) shows that observation density greatly affects the frequency of damaging hail reports, as well as whether or not damaging hail is observed at all on a given storm day. He states that a network comprised of one or more observation sites per square mile is necessary to adequately measure the areal extent of dam-

TABLE 12. Same as Table 6, but for 30 additional storm days.

Hail size (mm)	Number of observations	Percentage of observations less than model (%)	Average SHI (J m^{-1} s^{-1})
19–33	185	82	288
33–60	90	54	445
>60	39	8	609
All	314	65	373

aging hail. Since NWS spotter networks are much less dense than this, it is possible that many small-scale severe hail events, such as those produced by ''single-pulse'' type storms, are simply unreported. With single-pulse type storms relatively more common east of the Great Plains, this may be a significant factor leading to the higher algorithm FAR values in these regions.

Other potential causes of regional variation are differences in the typical storm environment. Numerical modeling studies have shown that the melting of hailstones is affected by a number of factors (Rasmussen and Heymsfield 1987). The most dominant factor is the thermal profile through which the hailstone falls. However, the RH profile also has a substantial effect. The HDA already incorporates some information on the vertical thermal profile (both the POH and POSH parameters are functions of the melting level). The impact that RH might have on the HDA was the focus of an additional study. Specifically investigated was whether the WTSM would benefit from the addition of an RH-dependent term (to its defining equation). Unfortunately, initial test results showed a minimal improvement in overall performance (the CSI increased by only 2%). However, for this study, the environmental RH was determined in a rather crude manner (using 700-mb upper-air plots), and so further investigation is needed to fully evaluate its effects.

Since the FAR is the statistic most variable on a regional basis, one might think that simply changing the WTSM to produce higher warning thresholds in those regions with higher FAR values would improve performance. Whereas this may be true in the MR region, that does not appear to be the case for the other regions (Figs. 10 and 11). Although the number of false alarms will decrease as the warning threshold increases, so too will the number of hits. And if, as the warning threshold is increased, the number of hits decreases more rapidly than the number of false alarms, the FAR will actually increase as the warning threshold rises. Therefore, despite the regional variations in overall CSI, it is not obvious that a separate WTSM is needed for each region. However, it is clear from the results shown in Fig. 13 that regional, or perhaps storm-environment-dependent, probability functions need to be developed. This will likely result in different optimum POSH thresholds (i.e., the threshold producing the highest overall CSI) for each region or storm environment, since the current model, optimized at 50%, does not appear to be appropriate for all regions or environments. And even within any one region, it will often not be best to always use just one overall, optimized threshold. For example, in situations where a storm is approaching a heavily populated area, a lower threshold may be better (given the results in Table 10).

It should be noted that all but two of the storm days used for the performance evaluation presented here came from WSR-88D sites at relatively low elevations ($<\sim$400 m) above mean sea level (MSL). Thus, the accuracy of the WTSM for WSR-88D sites at relatively high elevations ($\gtrsim$1 km) is questionable, since WT [as given by Eq. (4)] is a function of H_0 measured relative to the height ARL. Recent evaluation of HDA performance over Arizona (Maddox et al. 1998) indicates that, for the different WSR-88D sites located there, WT and POSH values can vary widely for constant melting level (relative to MSL) and SHI values, due solely to variations in the radar site elevation. There are also indications of a large overforecasting bias to POSH for high elevation WSR-88D sites (Kessinger and Brandes 1995; Maddox et al. 1998), due primarily to low values of WT for all seasons. Hence, until a terrain model can be added to the HDA and more extensive testing is done using data from high elevation WSR-88D sites, it may be necessary to change Eq. (4) so that H_0 is measured relative to MSL instead of ARL.

The prediction of maximum expected hail size is probably the most difficult and challenging aspect of the HDA. Also difficult is proper evaluation of the performance of the MEHS predictions, given the highly sporadic nature of the hail reports in *Storm Data.* The deficiencies in ground-truth verification of maximum hail size make scoring this component of the HDA very problematic. Without a high-density hail-observing network, one has no way of truly knowing the size of the largest hail being produced by a storm at any given time. The extent of this problem is amply illustrated by Morgan and Towery (1975). They present observations of a hailstorm on 21 May 1973, which moved across a very high-density (hailpads every 100–200 m) network located in Nebraska. Hail was observed at every site, with maximum sizes ranging from $\sim$1 to 3 cm. However, the area covered by the largest hail (3 cm) was only 1% of the total area of the network, with about 80% covered by hail $<$2 cm in diameter. Thus, at least in this case, the probability of a single hail report providing a true measure of the maximum hail size produced by the storm is very small. Therefore, due to the large uncertainties in the verification data, no attempt was made to determine any size-error statistics. Instead, only percentages of observed sizes greater than predicted sizes were calculated. Given the tendency for many different hail sizes to be observed for the larger SHI values, providing probabilities for various hail-size categories, in addition to a maximum size estimate, seems appropriate. And although correlations between the MEHS predictions and actual observed hail sizes will likely be poor at times, using the MEHS predictions as a relative indicator of overall hail damage potential may prove to be useful.

In addition to using SHI as a predictor of maximum hail size, Doppler-radar-determined storm-top divergence has been shown to be a reliable indicator of maximum hail size (Witt and Nelson 1991). A separate algorithm, called the upper-level divergence algorithm (ULDA), has been developed by the National Severe Storms Laboratory to detect and measure the strength

of these divergence signatures. During the early stages of HDA development, the ULDA was run concurrently with the HDA, and output from both algorithms was used to produce a final estimate of the maximum expected hail size. However, more extensive testing to date has shown that frequent problems associated with velocity dealiasing errors, range folding, and coarse vertical sampling when the WSR-88D is operating in VCP-21, degrade the ULDA's performance to the point that, overall, better size estimates are produced when solely using SHI as a predictor. It is hoped that future enhancements to the ULDA, along with better dealiasing techniques, will improve its performance to the point that it can make a positive contribution to the overall performance of the HDA. The detection and quantification of other radar signatures and/or environmental parameters may also help produce better MEHS predictions (e.g., bounded weak echo regions, midaltitude rotation, midaltitude winds).

At a minimum, the new WSR-88D HDA provides more information on the hail potential of a storm than the original hail algorithm. Test results also indicate that the new HDA outperforms the original hail algorithm. Steadham and Lee (1995) indicate that the original hail algorithm was not utilized much by operational warning forecasters. This may be due to poor performance and/or the limited nature of the output. With damaging hail being a significant hazardous weather threat, it is important that a hail detection algorithm produce guidance that a warning forecaster finds useful. Although additional improvements can certainly be made to the new HDA, its operational implementation will hopefully lead to more accurate and timely hazardous weather warnings.

Acknowledgments. We thank Robert Maddox, Conrad Ziegler and two anonymous reviewers for providing many useful comments that improved the manuscript. We also thank Joan O'Bannon for drafting two of the figures used in this paper. This work was partially supported by the WSR-88D Operational Support Facility and the Federal Aviation Administration.

REFERENCES

Brandes, E. A., D. S. Zrnic, G. E. Klazura, C. F. Suprenant, and D. Sirmans, 1991: The Next Generation Weather Radar (WSR-88D) as an applied research tool. Preprints, *25th Int. Conf. on Radar Meteorology,* Paris, France, Amer. Meteor. Soc., 47–50.

Browning, K. A., 1977: The structure and mechanisms of hailstorms. *Hail: A Review of Hail Science and Hail Suppression, Meteor. Monogr.,* No. 38, Amer. Meteor. Soc., 1–43.

Changnon, S. A., 1968: Effect of sampling density on areal extent of damaging hail. *J. Appl. Meteor.,* **7,** 518–521.

——, 1970: Hailstreaks. *J. Atmos. Sci.,* **27,** 109–125.

Crowther, H. G., and J. T. Halmstad, 1994: Severe local storm warning verification for 1993. NOAA Tech. Memo. NWS NSSFC-40, 30 pp. [NTIS PB 94215811.]

Crum, T. D., and R. L. Alberty, 1993: The WSR-88D and the WSR-88D Operational Support Facility. *Bull. Amer. Meteor. Soc.,* **74,** 1669–1687.

——, ——, and D. W. Burgess, 1993: Recording, archiving, and using WSR-88D data. *Bull. Amer. Meteor. Soc.,* **74,** 645–653.

English, M., 1973: Alberta hailstorms. Part II: Growth of large hail in the storm. *Alberta Hailstorms, Meteor. Monogr.,* No. 36, Amer. Meteor. Soc., 37–98.

Federer, B., and Coauthors, 1986: Main results of Grossversuch IV. *J. Climate Appl. Meteor.,* **25,** 917–957.

Foote, G. B., and C. A. Knight, 1979: Results of a randomized hail suppression experiment in northeast Colorado. Part I: Design and conduct of the experiment. *J. Appl. Meteor.,* **18,** 1526–1537.

Hales, J. E., Jr., and D. L. Kelly, 1985: The relationship between the collection of severe thunderstorm reports and warning verification. Preprints, *14th Conf. on Severe Local Storms,* Indianapolis, IN, Amer. Meteor. Soc., 13–16.

Johnson, J. T., P. L. MacKeen, A. Witt, E. D. Mitchell, G. J. Stumpf, M. D. Eilts, and K. W. Thomas, 1998: The Storm Cell Identification and Tracking (SCIT) algorithm: An enhanced WSR-88D algorithm. *Wea. Forecasting,* **13,** 263–276.

Kelly, D. L., J. T. Schaefer, and C. A. Doswell III, 1985: Climatology of nontornadic severe thunderstorm events in the United States. *Mon. Wea. Rev.,* **113,** 1997–2014.

Kessinger, C. J., and E. A. Brandes, 1995: A comparison of hail detection algorithms. Final report to the FAA, 52 pp. [Available from C. J. Kessinger, NCAR, P.O. Box 3000, Boulder, CO 80307.]

——, ——, and J. W. Smith, 1995: A comparison of the NEXRAD and NSSL hail detection algorithms. Preprints, *27th Conf. on Radar Meteorology,* Vail, CO, Amer. Meteor. Soc., 603–605.

Lemon, L. R., 1978: On the use of storm structure for hail identification. Preprints, *18th Conf. on Radar Meteorology,* Atlanta, GA, Amer. Meteor. Soc., 203–206.

Maddox, R. A., D. R. Bright, W. J. Meyer, and K. W. Howard, 1998: Evaluation of the WSR-88D Hail Algorithm over southeast Arizona. Preprints, *16th Conf. on Weather Analysis and Forecasting,* Phoenix, AZ, Amer. Meteor. Soc., 227–232.

Mather, G. K., D. Treddenick, and R. Parsons, 1976: An observed relationship between the height of the 45-dB*Z* contours in storm profiles and surface hail reports. *J. Appl. Meteor.,* **15,** 1336–1340.

Miller, L. J., J. D. Tuttle, and C. A. Knight, 1988: Airflow and hail growth in a severe northern High Plains supercell. *J. Atmos. Sci.,* **45,** 736–762.

Morgan, G. M., Jr., and N. G. Towery, 1975: Small-scale variability of hail and its significance for hail prevention experiments. *J. Appl. Meteor.,* **14,** 763–770.

NCDC, 1989: *Storm Data.* Vol. 31, No. 9, 58 pp.

NCDC, 1992: *Storm Data.* Vol. 34, No. 2–6.

Nelson, S. P., 1983: The influence of storm flow structure on hail growth. *J. Atmos. Sci.,* **40,** 1965–1983.

Petrocchi, P. J., 1982: Automatic detection of hail by radar. AFGL-TR-82-0277. Environmental Research Paper 796, Air Force Geophysics Laboratory, Hanscom AFB, MA, 33 pp. [Available from Air Force Geophysics Laboratory, Hanscom AFB, MA 01731.]

Polger, P. D., B. S. Goldsmith, R. C. Przywarty, and J. R. Bocchieri, 1994: National Weather Service warning performance based on the WSR-88D. *Bull. Amer. Meteor. Soc.,* **75,** 203–214.

Pratte, J. F., J. H. van Andel, D. G. Ferraro, R. W. Gagnon, S. M. Maher, and G. L. Blair, 1991: NCAR's Mile High meteorological radar. Preprints, *25th Int. Conf. on Radar Meteorology,* Paris, France, Amer. Meteor. Soc., 863–866.

Rasmussen, R. M., and A. J. Heymsfield, 1987: Melting and shedding of graupel and hail. Part II: Sensitivity study. *J. Atmos. Sci.,* **44,** 2764–2782.

Smart, J. R., and R. L. Alberty, 1985: The NEXRAD Hail Algorithm applied to Colorado thunderstorms. Preprints, *14th Conf. on Severe Local Storms,* Indianapolis, IN, Amer. Meteor. Soc., 244–247.

Steadham, R., and R. R. Lee, 1995: Perceptions of the WSR-88D performance. Preprints, *27th Conf. on Radar Meteorology,* Vail, CO, Amer. Meteor. Soc., 173–175.

Waldvogel, A., W. Schmid, and B. Federer, 1978a: The kinetic energy of hailfalls. Part I: Hailstone spectra. *J. Appl. Meteor.,* **17,** 515–520.

——, B. Federer, W. Schmid, and J. F. Mezeix, 1978b: The kinetic energy of hailfalls. Part II: Radar and hailpads. *J. Appl. Meteor.,* **17,** 1680–1693.

——, ——, and P. Grimm, 1979: Criteria for the detection of hail cells. *J. Appl. Meteor.,* **18,** 1521–1525.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

Winston, H. A., 1988: A comparison of three radar-based severe-storm-detection algorithms on Colorado High Plains thunderstorms. *Wea. Forecasting,* **3,** 131–140.

——, and L. J. Ruthi, 1986: Evaluation of RADAP II severe-storm-detection algorithms. *Bull. Amer. Meteor. Soc.,* **67,** 145–150.

Witt, A., and S. P. Nelson, 1991: The use of single-Doppler radar for estimating maximum hailstone size. *J. Appl. Meteor.,* **30,** 425–431.

——, M. D. Eilts, G. J. Stumpf, E. D. Mitchell, J. T. Johnson, and K. W. Thomas, 1998: Evaluating the performance of WSR-88D severe storm detection algorithms. *Wea. Forecasting,* **13,** 513–518.

Wyatt, A., and A. Witt, 1997: The effect of population density on ground-truth verification of reports used to score a hail detection algorithm. Preprints, *28th Conf. on Radar Meteorology,* Austin, TX, Amer. Meteor. Soc., 368–369.