Verification of Mesoscale Forecasts Made during MAP '88 and MAP '89

DING JINCAI,* CHARLES A. DOSWELL III, AND DONALD W. BURGESS†

NOAA, Environmental Research Laboratories, National Severe Storms Laboratory, Norman, Oklahoma

MICHAEL P. FOSTER AND MICHAEL L. BRANICK

NOAA, National Weather Service, Forecast Office, Norman, Oklahoma

(Manuscript received 25 October 1991, in final form 5 May 1992)

ABSTRACT

Two experimental forecasting projects, each called Mesoscale Applications Project (MAP), were conducted jointly by the National Severe Storms Laboratory and the National Weather Service Forecast Office at Norman, Oklahoma, during 1988 and 1989. This paper focuses primarily on the verification of the MAP '88 and MAP '89 experimental forecasts, and combines the results with those from a similar experiment run in 1987, to examine the evolution of forecast skill over that three-year period.

Results suggest that the severe-weather outlooks issued on a given experiment day exhibited good skill, with the skill being fairly stable over the three-year period studied. For outlooks issued the day before, the skill was notably higher in 1987 than in the subsequent years. Convective-mode forecasts ranged from poor to moderate skill levels, and did not change significantly from results obtained in 1987. Areal lightning forecasts were attempted in 1988 and 1989, with skill increasing more or less as the valid area increased, that area being defined as a circle ranging from 10 to 40 km of Norman. Advance outlooks for lightning, issued the day before the anticipated event, showed little or no skill. Some discussion of the possible reasons for the observed forecasting skill and its trends is presented. Several aspects of forecasting experiments in general are discussed also, based on experience during the MAP experiments.

1. Introduction

During the springs of 1988 and 1989, the National Severe Storms Laboratory (NSSL) and the National Weather Service (NWS) Forecast Office at Norman, Oklahoma, (Norman NWSFO) conducted a joint operational forecasting experiment called the Mesoscale Applications Project (MAP). As planned, MAP had two goals: 1) to provide data for basic and applied research on the issues of forecasting mesoscale/convective processes, and 2) to foster the interaction between researchers at NSSL and operational forecasters of the Norman NWSFO. These two experiments followed a similar experiment conducted during 1987 [called DOPLIGHT '87; see Doswell and Flueck 1989 (hereafter referred to as DF89) for some of the details of DOPLIGHT '87].

Forecast verification is a critical task in such a project since forecasting was the primary component of the experiments (see the discussion in DF89). This paper focuses on the verification of forecast products from the two experiments, MAP '88 and MAP '89. The experiment design is presented in section 2, including a description of the forecast products and a discussion of observations against which the forecasts were verified. Section 3 discusses verification methods, with the intent being to illustrate the merits and demerits of some typical verification measures that were used. Results from the experiments, including an analysis of how forecasting skill has evolved over the three years, are presented in section 4. A summary and discussion in section 5 concludes the paper.

2. Design of MAP

MAP '88 ran from 15 March to 24 June 1988, while MAP '89 ran from 6 March to 24 July 1989, for a total of 243 days during the two years. During the experiment, the NWSFO was staffed daily with a team of two extra "mesoscale forecasters," one from the NWSFO and one from NSSL. Team tasks included the following.

1. Routine preparation and dissemination of all MAP forecast products.

* On leave from Shanghai Meteorological Center, People's Republic of China.

† Current affiliation: NOAA/National Weather Service, WSR-88D Operational Support Facility, Norman, OK.

Corresponding author address: Dr. Charles A. Doswell III, National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069.

TABLE 1. Summary of forecast products; times denoted by "L" are local time (standard or daylight) and by "Z" are UTC. Note that for product 9, "24 h from 0000Z next day" is intended to account for the fact that the "date" changes at 0000Z.

Forecast product(s)	Issue time	Valid time	Dissemination
1. Mesoscale forecast discussion (MFD)	1645Z	1645–1200Z next day	AFOS only
2. Noon Outlook	1200L	1200L–1200Z next day	Local only
2.1 Severe—categorical			
2.2 Severe—probability			
2.3 Mesocyclone—categorical			
2.4 Mesocyclone—probability			
2.5 CG lightning—categorical			
2.6 CG lightning—probability			
3. Oklahoma thunderstorm outlook (OTO)	1300L	1300L–1200Z next day	Public
4. Graphic OTO	1300L	1300L–1200Z next day	Local only
5. Significant severe	1300L	1300L–1200Z next day	Local only
6. Nonsevere thunderstorm	1300L	1300L–1200Z next day	Local only
7. Nowcasts	as needed	2–3 h	Public
8. Convective mode	1500L	0000Z, 0600Z	Local only
9. Advance Outlook	1500Z–0000L	24 h from 0000Z next day	Local only
9.1 Severe—categorical			
9.2 Severe—probability			
9.3 CG lightning–categorical			
9.4 CG lightning—probability			

2. Running and evaluating applications programs. The programs included: quasigeostrophic diagnostics (Barnes 1985), AFOS (Automation of Field Operations and Services) Data Analysis Programs (ADAP, see Bothwell 1988), upper-air diagnostics (Foster 1988), AFOS cross-sectional and isentropic analyses, and hodograph analyses.

3. Maintenance of the MAP data archive, including real-time editing of AFOS surface hourly and upper-air base data, and execution of ADAP hourly.

4. Initiation of briefings with NWSFO public and aviation forecasters, as well as warning forecasters (when necessary), with an emphasis on mesoscale evolutions.

5. Collection of verification data for forecasts.

Table 1 provides a listing of the forecast products, not all of which will be discussed in this paper.[1] Product 1 is a technical discussion of the meteorological factors related to the day's forecast. It is an informal message composed by the MAP forecasters, with no prespecified details about format or style, to provide documentation of forecast reasoning for the day's products. Product 3, the Oklahoma Thunderstorm Outlook (or OTO), is a regularly scheduled release of the NWSFO during the convective season, so its text is more closely controlled, without technical language, and is issued to the public. The OTO was valid for the entire state of Oklahoma during MAP '88 and '89. Product 7 is a brief diagnosis and short-term forecast, again without technical terminology, that is released as needed by the NWSFO. These three products, being essentially narrative discussions, were not designed for objective verification and so are not considered here.

Products 4, 5, and 6 are simply graphical versions of the forecast. The forms for these were provided to the MAP forecasters, to include a base map of Oklahoma on which the outlined area(s) (if any) were drawn. Since these graphical products merely provide visual documentation of the forecasts, they are not considered in this paper, either.

The Noon Outlook (product #2) is valid for that portion of Oklahoma within 230 km (roughly, 125 n mi) of the NSSL Norman Doppler radar (see Fig. 1a). As in DOPLIGHT '87, both categorical (or dichotomous) and probability (or polychotomous) forecasts were issued. The intent in issuing both types of forecasts was to collect evidence about how human forecasters go about the subjective thresholding process. Therefore, no guidance was offered to the forecasters about how to reconcile the categorical and probabilistic forecasts; as noted by Murphy (1991b), consistent conversion of probability forecasts into categorical forecasts can be mandated in advance, and a threshold probability of 50% need not be assumed.

In Doswell et al. (1990, hereafter referred to as DDK90), results suggested that forecasters are not necessarily "consistent" in their choice of thresholds, in the sense that when their polychotomous forecasts were verified with methods comparable to those used for dichotomous forecasts, the results were not the same. Put another way, different forecasters may have used different thresholds when converting between categorical and probabilistic forecasts, and a given forecaster may have used different thresholds on dif-

[1] The full set of MAP forecast products is included simply to point out that the team had many responsibilities besides those that can be discussed in a paper about verification, as was the case in the DOPLIGHT '87 experiment.

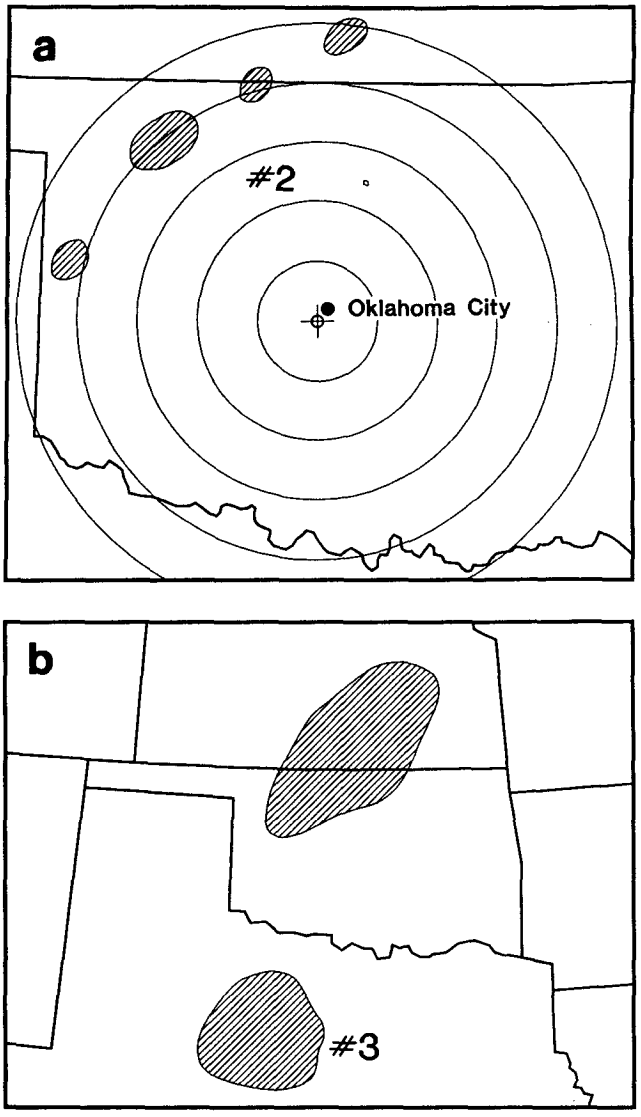

FIG. 1. Map showing forecast verification areas for (a) radar-based convective-mode forecasts and (b) satellite-image–based convective-mode forecasts. Also shown are schematic examples of each type of forecast. In (a) mode #2 shows a group of isolated cells (scattered along a line), while in (b) mode #3 shows two meso-β-scale convective systems. While the choice of convective mode in (a) is arguable, this illustrates some of the problems associated with defining the convective mode. In (a), while there are multiple cells and they are scattered along a line, they are separated enough that the *dominant* character of the radar echoes is their relatively isolated character.

ferent days. The degree of inconsistency revealed in DDK90 (seen in their Fig. 3) is not clearly indicative of a major problem, however. Moreover, nonfixed thresholds can even be considered desirable. Such variable thresholds are being used daily in an objective procedure that generates dichotomous convective outlooks for thunderstorms and severe local storms (described in Reap et al. 1982). It appears that, even without explicit guidance about how to go about it, human forecasters are capable of converting probabilities into dichotomous forecasts without creating notable problems. It was decided to continue this form of experiment during MAP '88 and '89.

It can be seen in Table 1 that the issue and valid times for the various products are a mixture of UTC and local times. This is an undesirable but inescapable consequence of doing the experiment in an operational forecasting office. We have defined the times in accordance with operational exigencies and provide them here for completeness.

For the dichotomous and polychotomous forecasts related to severe weather,[2] observed events for verification were determined by whether or not one or more warnings were issued by the warning forecaster from the NWSFO staff, in an effort to avoid problems arising from uneven reporting of severe weather (as discussed in DF89). Whether or not this approach solves the problems associated with severe-event reporting certainly is subject to debate. We make no pretense that there are no problems with this approach; it is simply a choice we made.

In view of the fact that the MAP forecasters interacted with the NWSFO staff each day, the warnings issued by the Norman NWSFO cannot be said to be completely independent of the MAP forecasts. However, the NWSFO forecasters had no vested interest in the success of the MAP forecasts; rather, the NWSFO forecasters (as always) simply were trying to do the best possible job of warning verification and had nothing to gain from issuing warnings to improve MAP verification.

Since the valid areas for severe weather forecasts issued by the MAP forecasters included regions for which warnings are issued by NWS offices other than NWSFO Norman, warnings issued by those other offices (including Tulsa, Oklahoma; Fort Smith, Arizona; Amarillo, Texas; and Forth Worth, Texas) were used for verification in the appropriate situations. Mesocyclone presence was determined by using the Doppler radar; if the formal criteria for a mesocyclone were met (see Burgess and Lemon 1990) then a mesocyclone had been observed for verification purposes.

Lightning forecasts were for cloud-to-ground (CG) strikes within a given distance (initially set within a radius of 20 km) from Norman, and the verifying data were taken from the NSSL lightning-detection network (see Mach et al. 1986) datasets. During the verification, some adjustments to this were tried, as shall be discussed below.

The Advance Outlook (product #9) in MAP '88 is essentially identical to the Noon Outlook product, except that it is issued the day before. Verification criteria

[2] The definition of severe weather follows standard National Weather Service practice: damaging winds (50 kt or greater), large hail (3/4 in. or greater), and tornadoes.

and data are also the same as the Noon Outlook. In previous experiments (e.g., DOPLIGHT '87 as described in DF89), the Advance Outlook was issued in support of field research observations. During the planning for MAP '89, it was determined that there was no need for this product in the absence of a special field-observing program, so the Advance Outlook was eliminated in MAP '89. Moreover, operational forecasters do not have a specific requirement for such a long-term outlook, so it has not served any particular operational need. The absence of a formal Advance Outlook product does not mean that the NWSFO and MAP forecasters did not interact *informally* about the possibility of severe weather in that time period.

As in DOPLIGHT '87, an attempt was made during MAP '88 to forecast the so-called "convective mode"—that is, the "dominant" structures seen in the day's convection (product #8). There is considerable anecdotal evidence that the convective mode influences the sorts of weather phenomena that will occur on a given day, and it seems desirable to be able to forecast convective mode. Unfortunately, the anecdotal evidence is not documented very well in the literature. From our perspective, it seems that a reason for this lack of documentary evidence is that convective mode turns out to be rather difficult to define in a unique, objective, and verifiable way (see DF89).[3] Since DF89 pointed out some flaws in the way DOPLIGHT '87 handled the convective-mode forecast verification, an attempt was made in MAP '88 to improve on the product design. Unfortunately, it continued to prove troublesome and was eliminated during MAP '89.

The convective-mode experiment in MAP '88 defined convective mode in two different ways: radar based and satellite-image based (see Table 2). In order to circumvent the problems encountered during DOPLIGHT '87, the forecasts were valid precisely at 0000 and 0600 UTC, rather than over an extended period as in DOPLIGHT '87. The radar-based mode forecasts were defined over an area within 230 km of the Oklahoma City WSR-57 radar; the forecaster was to outline the expected echoes on the base map (Fig. 1a) without trying to delineate reflectivity contours, and choosing a mode from the appropriate part of Table 2.

The satellite-based mode forecasts were over a larger area (Fig. 1b), but the procedure was similar. That is, the forecaster was to outline the expected areas of convective clouds (cold topped) as depicted by infrared (IR) satellite images and pick a mode from Table 2. As with the radar-based convective mode, forecasts of the cloud-top equivalent blackbody temperature contours were not required. The Oklahoma City WSR-57 radar films and appropriate IR satellite images were used for verification.

[3] We shall provide additional discussion on this issue later in the paper.

TABLE 2. MAP '88 convective forecast modes. The so-called "Maddox criteria" are those set forth in Maddox (1980) for mesoscale convective complexes (MCCs), and those convective sytems that fail to meet the Maddox criteria are designated meso-α and meso-β, according to the scale definitions described in Orlanski (1975).

Mode (radar)	Description	Mode (satellite)	Description
0	no convective echo	0	no convective clouds
1	isolated cells	1	MCC (Maddox criteria)
2	multicell echoes		
3	supercell echoes	2	meso-α convective system
4	squall line(s)		
5	other	3	meso-β convective system
		4	squall line(s)
		5	isolated convective clouds
		6	other

3. Verification methods

As in DOPLIGHT '87, both categorical and probabilistic forecasts were used. Some standard verification measures have been employed herein for categorical forecast verification using the 2×2 contingency table. These include probability of detection (POD), false-alarm ratio (FAR), critical success index (CSI), probability of false detection (POFD), true skill statistic (TSS), and Heidke skill score (S), as described in Donaldson et al. (1975), DF89, and DDK90. The POD is the ratio of the number of events correctly forecast to the total number of observed events; FAR is the ratio of the number of events forecast incorrectly to the total number of forecast events; POFD is the ratio of the number of events forecast incorrectly to the total number of observed nonevents; CSI is the ratio of the number of events correctly forecast to the sum of observed events and the number of incorrectly forecast events. The CSI ignores those nonevents that were correctly forecast, whereas the TSS and S attempt to account for correctly forecast nonevents. The TSS and S scores both compare the accuracy of forecasts to a standard "random" forecast; the TSS can be expressed simply as the POD minus the POFD, while the S score uses the "random" forecast standard in a subtly different way. Each of these three has its own advantages and disadvantages, as discussed in the references.

We note, however, that the CSI and S values have an advantage over POD and TSS, in that blindly maximizing the latter scores may lead one to an overforecasting strategy (see DF89 and DDK90). When dealing with rare events, POFD typically remains small, since nonevents dominate rare-event forecasting and the number of nonevents is in the denominator of the POFD.

As mentioned briefly in DDK90, the off-diagonal elements of the 2×2 contingency table can be interchanged without affecting either CSI or S. That is, there

is nothing in either the CSI or S to distinguish between over- and underforecasting. Of course, a bias can be calculated to determine this. However, consideration of the POD and FAR can allow one to distinguish clearly between over- and underforecasting. As noted in DDK90, there are good reasons why no single number can represent clearly the results of a verification (see also the exchange between Murphy 1991b; and Doswell et al. 1991).

When considering probability forecasts, it has been shown in DDK90 that the same verification measures as in categorical forecasting can be employed. However, it is useful to add the half-Brier Score (*Bf*, Brier 1950), the skill score based on Brier score (*Bs*, Sanders 1963), and the bias (see Hughes 1980). In order to determine these scores, we use the following quantities: Ro is the observed frequency of the event during the experiment, Rf is the average of the forecast probability, and *Bc* is the half-Brier score for a constant climatology (or, alternatively, the observed frequency) forecast; here, we have used the observed frequency to find *Bc*.

The *Bf* score measures the error between forecast probabilities and observed frequencies; thus, the smaller it is, the better the forecasts. *Bs* is a skill score comparing the forecast *Bf* score to that derived using the constant "climatological" probability forecast; specifically, $Bs = 100\,(Bc - Bf)/Bc$. Thus, the larger *Bs* is, the better the forecast.

For rare-event forecasts, the low-probability cases dominate, so *Bf* varies only slightly (at a relatively low level) for convective forecasts. In other words, *Bf* is not very sensitive to severe convective forecasting skill owing to severe convection being a rare event. Thus, *Bs* and bias will be the primary measures of skill for probability forecasts.

It should be pointed out that the preceding forecast verification methodology typically is applied to point forecasts, whereas we have used it to verify forecasts over relatively large areas. There was no attempt to "pinpoint" the severe-weather events *within* the valid areas, as such an effort was considered well beyond the current capabilities of human forecasters. In fact, an attempt to forecast the county and hour of the first convection of the day was done during MAP '88. The forecasts were so poor that we have chosen not to present them; at least one reason for this choice was that it proved difficult to define what was "first convection" (as it has proven difficult to define "convective mode"). What criteria would distinguish between a cumulus cloud and a "convective storm"? Radar, CG lightning, and satellite criteria for doing so all have problems that made this more difficult than we first thought. Even accounting for these difficulties, it appears that "pinpointing" convective events in space and time will be a considerable challenge!

TABLE 3. Observed weather events (as defined in each experiment) for DOPLIGHT '87, MAP '88, and MAP '89; frequencies relative to the total are shown in parentheses.

	Occurrence days	Mesocyclone days	Total
DOPLIGHT '87	26 (29%)	6 (7%)	91
MAP '88	19 (19%)	3 (3%)	102
MAP '89	49 (35%)	17 (12%)	141
Total	94 (28%)	26 (8%)	334

TABLE 4. Combined MAP '88 and MAP '89 contingency table and summary statistics for noon severe-weather outlook.

	Observed		
Forecast	Yes	No	Total
yes	49	19	68
no	19	155	174
total	68	174	242

POD = 0.72 FAR = 0.28 POFD = 0.11 CSI = 0.56
TSS = 0.61 S = 0.61

One way to approach the problem of verifying area forecasts would be to issue purely probabilistic forecasts (as advocated recently by Murphy 1991a); for example, in the form of probability contours over the forecast domain [as is done, for example, in the objective thunderstorm and severe local-storm forecasting technique developed by Reap and Foster (1979)]. A similar approach might be to break down the forecast area into a number of smaller subdomains and issue separate forecasts (dichotomous or polychotomous) for each subdomain. Moreover, our forecasts were for *one or more* events within the forecast areas, making no distinction between days with numerous events and those with only one, and ignoring any differences in the *intensity* of the events. While experiments incorporating these sorts of refinement ought to be done in the future, we simply did not do so in either of the MAP experiments.

4. Verification results

The 1988 spring season in Oklahoma was characterized, as in 1987, by a relative scarcity of severe-weather events. This implies that the small sample size for some of the forecast products (notably, mesocyclones) continues to be a problem for statistical reliability. In contrast, 1989 was more typical of Oklahoma springs, with many more events. Table 3 shows the number of observed events during the three years.

a. Severe-weather outlooks

For the Noon Outlooks, the verification results for the combined seasons of 1988 and 1989 are shown in Tables 4 and 5. Since the observed data are missing from 6 May 1989, the sample size is 242 (one less than the total number of operating days). As delineated in

TABLE 5. Combined MAP '88 and MAP '89 noon severe-weather outlook probability forecast evaluation table. Statistics are derived using the given probability category as a threshold for converting to dichotomous forecasts (see DF89).

Probability category	Number of forecasts	Hits	POD	FAR	CSI	POFD	TSS	S
1.00	5	5	.07	.00	.07	.00	.07	.10
.90–.99	10	8	.19	.13	.19	.01	.18	.24
.80–.89	7	6	.28	.14	.27	.02	.26	.33
.70–.79	7	6	.37	.14	.35	.02	.34	.42
.60–.69	10	7	.47	.18	.43	.04	.43	.49
.50–.59	19	13	.66	.22	.56	.07	.59	.61
.40–.49	17	5	.74	.33	.54	.14	.59	.57
.30–.39	12	4	.79	.38	.53	.19	.60	.56
.20–.29	16	6	.88	.42	.54	.25	.64	.55
.10–.19	29	4	.94	.52	.47	.39	.55	.43
.01–.09	58	3	.99	.65	.35	.71	.28	.18
0.0	52	1	1.0	.72	.28	1.0	.00	.00

Bf = 0.12 Ro = 0.28 Rf = 0.27 Bc = 0.20 Bs = 42.52 bias = −3.31

Table 4, the CSI and S values of 0.56 and 0.61, respectively, indicate good accuracy and are only slightly smaller than the comparable results from DOPLIGHT '87 (0.61 and 0.67).[4] This indicates reasonable stability over the three-year test period, with the following values averaged over all three years: POD = 0.72, FAR = 0.26, POFD = 0.10, CSI = 0.58, TSS = 0.62, and S = 0.62. Using a χ^2 test, the statistical confidence in the results from Table 4 far exceeds the 99.95% level (χ^2 of 90.46, versus a 99.95% level of 12.12). It is perhaps only coincidental that the missed events exactly equal the false alarms (19), showing zero bias.

As shown in Table 5, the value of *Bs* (42.52) indicates about a 40% improvement over the "climatological" forecasts by the probabilistic version of the Noon Outlook. The bias is negative but quite small (−3.31). These results, again, are quite comparable to those shown in DF89 for DOPLIGHT '87. For the combined three years, we have the following results: *Bf* = 0.12, *Bs* = 43.11, bias = −3.51.

It is interesting to note that the maximum S value (0.61) in Table 5 suggests that the best choice of a threshold probability category for converting probabilistic forecasts into categorical forecasts[5] is the probability category 0.50–0.59 (see DF89). Maximum S also corresponds to a maximum value for the CSI. Using this as a threshold, the contingency-table–based parameters (e.g., POD, FAR, POFD, CSI, etc.) are quite close to those derived from the actual categorical forecasts (compare Table 5, line 6 with Table 4). If the TSS is maximized when choosing a threshold (as suggested in DF89), however, a quite different probability category would be chosen (0.29–0.20), in this case exhibiting an associated overforecasting tendency for the categorical forecasts converted from the probabilistic forecasts using the TSS-based threshold. It turns out that this was not true for the results shown in DF89; in 1987, the threshold probability category was the same whether one maximized TSS or S. Based on our experience, it appears that thresholding using S is preferable to that using TSS, because in our experiments the TSS has a marked tendency to peak in a lower probability category than S. This leads to overforecasting when that category is used as a threshold, paralleling the concerns expressed in DDK90 to the effect that the TSS tends to be overly influenced by the POD. However, at this point, it cannot be asserted that this is a *general* result; we have focused on rare events and it may well be that for common (i.e., not rare) events, a TSS-based threshold would give satisfactory results.

[4] The definition of a severe-weather event in DOPLIGHT '87 differs somewhat from that used in MAP '88 and '89. In 1987, an event was determined by the DOPLIGHT Doppler Radar Interpreter (as described in DF89). Thus, there might be some concern for a direct comparison between results; although there is no easy way to prove it, we believe the practical differences are not critical and the comparison is probably valid.

[5] The sense of what is the "best" choice depends on one's intent, however. There may well be situations wherein one might accept a high FAR in order to obtain a high POD. This is discussed at some length in Murphy (1991a) in terms of biased dichotomous forecasts. We agree with Murphy that biased forecasting is not desirable, in principle.

TABLE 6. MAP '88 contingency table and summary statistics for advance severe-weather outlook.

	Observed		
Forecast	yes	no	total
yes	10	8	18
no	10	74	84
total	20	82	102

POD = 0.50 FAR = 0.44 POFD = 0.10 CSI = 0.36
TSS = 0.40 S = 0.42

TABLE 7. MAP '88 advance severe-weather outlook probability forecast evaluation table, as in Table 5.

Probability category	Number of forecasts	Hits	POD	FAR	CSI	POFD	TSS	S
1.00	0	0	.00	.00	.00	.00	.00	.00
.90–.99	2	2	.10	.00	.10	.00	.10	.15
.80–.89	0	0	.10	.00	.10	.00	.10	.15
.70–.79	0	0	.10	.00	.10	.00	.10	.15
.60–.69	5	3	.25	.29	.23	.02	.23	.30
.50–.59	7	4	.45	.36	.36	.06	.39	.44
.40–.49	5	1	.50	.47	.34	.11	.39	.40
.30–.39	4	1	.55	.52	.34	.15	.40	.38
.20–.29	8	2	.65	.58	.34	.22	.43	.36
.10–.19	25	5	.90	.68	.31	.46	.44	.26
.01–.09	32	1	.95	.78	.21	.84	.11	.05
0.0	14	1	1.0	.80	.20	1.0	.00	.00

Bf = 0.12 Ro = 0.20 Rf = 0.20 Bc = 0.16 Bs = 23.00 bias = 0.00

The Advance Outlook product was omitted in 1989, so verification results are available only for MAP '88. Table 6 presents the dichotomous forecasts, while Table 7 shows the polychotomous results. If one compares these to the comparable forecasts for DOPLIGHT '87 (shown in DF89), it appears that the skill level for Advance Outlooks has decreased. In trying to understand the origins of this apparent change in skill, we noted that Advance Outlook forecasts are based mainly on the operational numerical weather prediction (NWP) model forecasts, specifically the 24–48-h products. This raises an issue about our forecasts; namely, the role of the NWP model forecasts in influencing the human-generated forecasts. In order to examine the general quality of the NWP forecasts during MAP '88, a separate verification of selected nested-grid model (NGM) 24-h forecast products was conducted. Note that all nonevent days (irrespective of the forecasts) were excluded from consideration for this evaluation.

This verification consisted of establishing error categories (Table 8) and determining the frequency of NWP forecasts in those categories. The area over which the forecasts were verified was limited to Oklahoma and its neighboring states.

As shown in Table 9, more than 80% of the situations during MAP '88 were within two categories of being correct. This indicates that the NWP model guidance was reasonably good during the experiment, and thus is not a likely source for the degradation in forecast skill compared to DOPLIGHT '87. As of this writing, the reasons for this decline in Advance Outlook forecast skill in 1988 remain unknown. In any case, it is not likely that two years of Advance Outlook forecast evaluation is going to yield conclusive results, but we hope eventually to find at least a plausible hypothesis for future testing.

b. Mesocyclone Noon Outlook

During the 102 forecast days of MAP '88, there were only three mesocyclone days (i.e., days on which one or more mesocyclones were observed), all of which were not forecast. Conversely, there were three days on which mesocyclones were forecast, none of which verified. When these are combined with the more frequent events of 1989, Table 10 results. These results indicate relatively poor skill at mesocyclone forecasting, and are even worse than the limited results of DOPLIGHT '87 (see DF89).

Table 11 shows that the mesocyclone probability forecasts exhibited modest skill when compared to the "climatology" forecasts for mesocyclones, but still well below the skill levels exhibited for severe-weather events in general. The relatively high bias certainly indicates an overforecasting tendency. Note that the small mesocyclone Noon Outlook *Bf* value (0.06) is even smaller than that associated with the general Noon Outlook product (0.12). This does not imply that the mesocyclone forecasts were better, however, as suggested by the *Bs* values (25.72 vs 42.52). The main cause for the low *Bf* scores is the low-event frequency, and correct forecasts of nonevents assign a value of zero to *Bf*, suggesting that *Bf* alone is not a good mea-

TABLE 8. NGM 24-h forecast products tested for accuracy during MAP '88; all errors are determined as forecast minus observed. The error category is found by rounding the observed error as measured; e.g., a temperature error of −3.5 deg C belongs to the −4 category, whereas an error of +2.3 deg C belongs to the +2 category. System location refers to the position of features like low or high pressure centers, fronts, and troughs or ridges. 850-hPa temperature is a subjectively estimated area average temperature in Oklahoma. Other values are maxima in the forecast area.

Product	Category increment
1. Surface system location	1 deg latitude (at 40°N)
2. 850-hPa system location	1 deg latitude (at 40°N)
3. 500-hPa system location	1 deg latitude (at 40°N)
4. 850-hPa temperature	1 deg C
5. Surface–500-hPa rel. humidity	10%
6. 700-hPa vertical velocity	10^{-3} hPa s^{-1}
7. 500-hPa vorticity	10^{-5} s^{-1}

TABLE 9. Results of verification of products listed in Table 8, shown as frequencies (%) in error categories, with the mean category error.

Product	≤ −4	−3	−2	−1	0	+1	+2	+3	+4	Mean
1	0	0	0	2	92	4	0	2	0	0.09
2	0	0	0	4	81	9	6	0	0	0.17
3	0	0	2	9	83	2	4	0	0	−0.02
4	0	0	2	0	70	0	20	4	4	0.68
5	0	0	0	2	49	28	19	2	0	0.70
6	0	2	4	2	58	9	4	15	6	0.81
7	11	0	13	0	48	2	11	0	0	−0.85

sure of forecast skill when rare events are considered. Hence, its use is discontinued in what follows.

c. *Convective-mode forecasts*

The radar- and satellite-image–based convective-mode forecasts were discontinued in 1989, so results are confined to MAP '88. Tables 12–15 show the verification results for the 0000 and 0600 UTC convective mode forecasts. The highest accuracy is associated with the 0000 UTC satellite-image–based mode product, with S = 0.42 indicating only a moderately accurate forecasts. The relatively low S values are certainly no better than that exhibited during DOPLIGHT '87 (see DF89). These results confirm the findings of DF89 that only modest skill is shown at forecasting convective mode.

However, any conclusion must be tempered by the fact that large convective systems (meeting the Maddox 1980 criteria for an MCC) again avoided the area during MAP '88, as during 1987. There were only two such systems (on 31 March and 5 May) and they both were not forecast, with one false alarm (on 14 June). Thus, for the two years of convective-mode forecasts (1987 and 1988), meso-α-scale (see Orlanski 1975) convective events were rather uncharacteristically infrequent.

Deficiencies in convective mode forecast design became apparent after the experiment got underway, as in DOPLIGHT '87. During MAP '88, it was hoped that redesign of this part of the experiment would overcome some of the deficiencies noted in DF89. It appears, however, that the number of modes offered to the forecasters (cf. Table 2) was too high; some modes had fewer than five events during MAP '88, yielding small sample sizes for verification purposes.

Also, it appears that while forecasting for specific times simplifies the *verification,* it significantly complicates the *forecasting.* Forecasting the mode is difficult enough, because the mode changes during a typical convective evolution. By making the forecasts valid only at specific times, the forecaster had to forecast which mode the system would be in at the specified time. The results show that the radar-based forecasts were rather less accurate on the whole than the satellite-based forecasts. This may be the result of the relatively small area associated with the radar-based forecasts.

One hopeful indication from the convective-mode forecasts concerns mode 4: squall lines. It appears that this convective mode could, indeed, be anticipated, especially for the satellite-based forecasts.

TABLE 10. Combined MAP '88 and '89 contingency table and summary statistics for noon mesocyclone outlook.

	Observed		
Forecast	Yes	No	Total
yes	8	11	19
no	12	211	223
total	20	222	242

POD = 0.40 FAR = 0.58 POFD = 0.05 CSI = 0.26
TSS = 0.35 S = 0.36

e. *Lightning forecasts*

As with the other forecast elements, both dichotomous (categorical) and polychotomous (probability) forecasts were issued, in a Noon Outlook and Advance Outlook mode. The latter was canceled during MAP '89. Although the forecast design called for forecasts only of events within 20 km of Norman, the verification was done as if the forecasts were for events within other radii (10, 30, 40, and 50 km), as well. It was felt that this could indicate what might be a reasonable radius to impose on any future forecasting experiments, since 20 km had been chosen quite arbitrarily.

Figure 2 shows how the verification measures vary as a function of radius for the dichotomous noon lightning outlook. Both FAR and POFD decrease more or less linearly with increasing radius, whereas CSI and S values increase linearly with increasing radius, with some indication of the increase leveling off at 40 km and beyond. On the other hand, Fig. 3 shows that the Advance Outlook exhibits somewhat different behavior; the CSI and S skill scores show an overall decline with increasing radius. The possible reasons for these observed variations in skill with radius are not clear, as of this writing; we feel we have too little experience with this product to offer speculations at this time.

TABLE 11. Combined MAP '88 and MAP '89 mesocyclone Noon Outlook probability forecast evaluation table, as in Table 5.

Probability category	Number of forecasts	Hits	POD	FAR	CSI	POFD	TSS	S
1.00	0	0	.00	.00	.00	.00	.00	.00
.90–.99	1	1	.05	.00	.05	.00	.05	.09
.80–.89	0	0	.05	.00	.05	.00	.05	.09
.70–.79	2	2	.15	.00	.15	.00	.15	.24
.60–.69	2	2	.25	.00	.25	.00	.25	.38
.50–.59	5	2	.35	.30	.30	.01	.34	.44
.40–.49	13	3	.50	.57	.30	.06	.44	.41
.30–.39	9	2	.60	.63	.30	.09	.51	.40
.20–.29	23	4	.80	.71	.27	.18	.62	.35
.10–.19	26	3	.95	.77	.23	.28	.67	.28
.01–.09	72	0	.95	.88	.12	.61	.34	.09
0.0	88	1	1.0	.92	.08	1.0	.00	.00

Bf = 0.06 Ro = 0.08 Rf = 0.12 Bc = 0.08 Bs = 25.72 bias = 43.75

For the polychotomous forecasts, the Noon Outlook lightning *Bs* scores (Fig. 4) seem to indicate a nearly constant skill level at 30 km and beyond. On the other hand, the Advance Outlook lightning *Bs* scores indicate little or no skill at any radius. The bias values show that overforecasting is the rule within 20 km and underforecasting dominates at 30 km and larger radii.

5. Summary and discussion

Over the three years of forecasting experiments, the Noon Outlook severe-weather product skill seems stable and remains at a relatively high level. It can be surmised that this is the result of years of experience[6] in forecasting severe weather accumulated by the DOPLIGHT and MAP forecast teams.

On the other hand, the Advance Outlook for severe weather exhibited a decrease during MAP '88, for reasons that remain unclear. This can be interpreted as a note of caution about the findings in DF89, in which it was found that the Advance Outlook was nearly as skillful as the Noon Outlook. Some attempt will be made in the future to explore the causes for this difference; our findings in this report suggest that the source probably is not to be found in the NWP model guidance.

Taken in total, the three years' worth of experiments leave us with mixed results about forecasting the presence of mesocyclones. The sample size is still too small to indulge in sweeping generalizations, but it appears that human forecast skill in distinguishing potential mesocyclonic storms is relatively modest, at least in comparison to the prediction of severe storms in general. In spite of continuing expansion of our knowledge through numerical cloud simulations (e.g., Weisman and Klemp 1986), especially how mesocyclonic storms arise from convective interaction with its environment,

[6] This experience is mostly based on forecasting outside the experiments themselves.

TABLE 12. Radar-based convective-mode contingency table for 0000 UTC, including summary statistics.

	Observed					
Forecast	0	1	2	3	4	Total
0	57	1	2	0	0	60
1	19	8	4	1	3	35
2	1	6	2	0	1	9
3	0	0	0	0	0	0
4	4	0	1	0	3	8
Total	81	14	9	1	7	112
Statistic						
POD	.70	.57	.22	.00	.43	
FAR	.05	.77	.78	.00	.63	
CSI	.68	.20	.13	.00	.25	
POFD	.10	.28	.07	.00	.05	
TSS	.61	.30	.15	.00	.36	
S	.50	.18	.15	.00	.36	

Total TSS = 0.42 Total S = 0.33

TABLE 13. As in Table 12, except for 0600 UTC.

	Observed					
Forecast	0	1	2	3	4	Total
0	72	5	2	0	0	79
1	6	4	4	0	1	15
2	4	1	0	0	2	7
3	1	0	0	0	0	1
4	3	0	1	0	1	5
Total	86	10	7	0	4	107
Statistic						
POD	.84	.40	.00	.00	.25	
FAR	.09	.73	1.0	1.0	.80	
CSI	.77	.19	.00	.00	.13	
POFD	.33	.11	.07	.01	.04	
TSS	.50	.29	−.07	−.01	.21	
S	.45	.23	−.07	.00	.19	

Total TSS = 0.32 Total S = 0.28

TABLE 14. MAP '88 satellite-based convective-mode contingency table for 0000 UTC, including summary statistics.

	Observed							
Forecast	0	1	2	3	4	5	6	Total
0	28	1	0	4	0	3	0	36
1	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	1
3	4	0	1	19	1	10	0	35
4	2	0	0	0	7	0	0	9
5	13	1	2	12	7	35	0	70
6	1	0	0	0	0	0	0	1
Total	48	2	3	36	15	48	0	152
Statistic								
POD	.58	.00	.00	.53	.47	.73	.00	
FAR	.22	.00	1.0	.46	.22	.50	1.0	
CSI	.50	.00	.00	.37	.41	.42	.00	
POFD	.08	.00	.01	.14	.01	.34	.01	
TSS	.51	.00	−.01	.39	.45	.39	−.01	
S	.54	.00	−.01	.39	.55	.35	.00	

Total TSS = 0.42 Total S = 0.42

this apparently does not yet imply a concomitant increase in forecasting skill. This can be attributed to two sources of error: physically incorrect simulations, and an inability to forecast the environment. We are not prepared in this report to go into the question of assigning forecast errors to these two distinct sources, but we think our experiments probably have established a baseline measure of current subjective mesocyclone forecasting skill.

Although plagued by continuing problems in experimental design with regard to convective-mode forecasts, it appears that DOPLIGHT '87 and MAP '88 at least have served to refine how such experiments might be conducted in the future. It appears that forecasters show reasonable skill in anticipating convective versus nonconvective days, but that differentiating among different convective modes, given that convection occurs, is a daunting task. Apparently, there may be some skill in distinguishing "squall-line" from "non-squall-line" days, but such a conclusion can be only tentative, at best. If such skill exists, as of this writing we do not have a clear picture of the factors upon which it rests. A continuing problem for experimental design is that convection typically undergoes mode changes during a convective episode, and part of the problem with forecasting convective mode is assigning a time period to a particular mode. Moreover, different sensors (e.g., radars and satellite imagery) observe convection differently, making mode assignment sensor dependent. Assigning modes to observations remains a highly subjective process, further complicating the experiments. Obviously, this is an area ripe for further research.

Regarding the cloud-to-ground (CG) lightning forecasts, it appears that making an Advance Outlook (i.e.,

TABLE 15. As in Table 14, except for 0600 UTC. Mode 6 was neither forecast nor observed, so it has been omitted from the table.

	Observed						
Forecast	0	1	2	3	4	5	Total
0	46	1	2	10	0	2	61
1	0	0	0	1	0	0	1
2	1	0	0	3	0	0	4
3	12	0	0	12	1	5	30
4	2	0	0	0	5	0	7
5	4	0	2	5	1	8	20
Total	65	1	4	31	7	15	123
Statistic							
POD	.71	.00	.00	.39	.71	.53	
FAR	.25	1.0	1.0	.60	.29	.60	
CSI	.57	.00	.00	.24	.56	.30	
POFD	.26	.01	.03	.20	.02	.11	
TSS	.45	−.01	−.03	.19	.70	.42	
S	.45	−.01	−.03	.19	.70	.37	

Total TSS = 0.36 Total S = 0.35

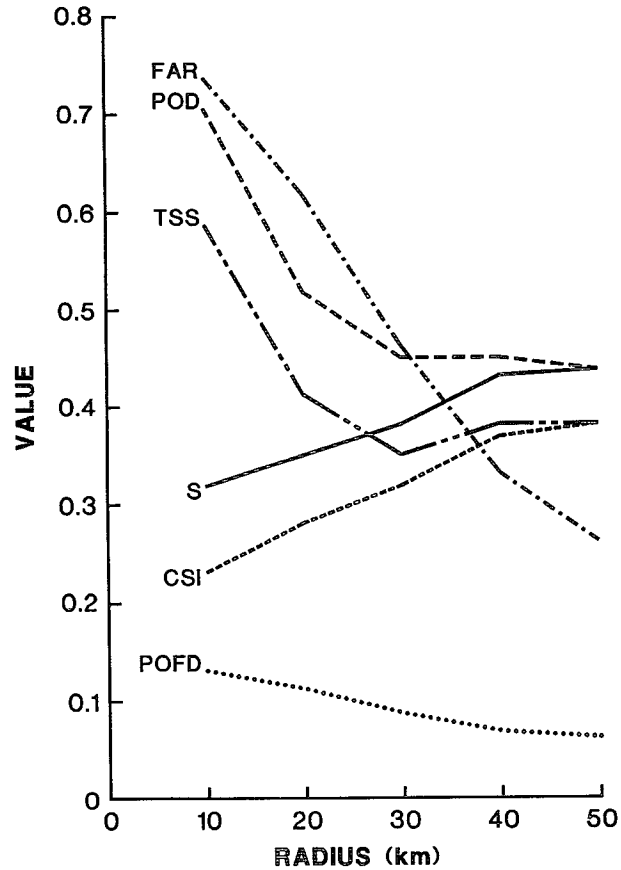

FIG. 2. Verification scores for the Noon Outlook lightning categorical forecasts as a function of the radius.

24–36-h forecasts) of CG lightning for a specific area is unlikely to show much skill. However, on the day of the expected event, it may be possible to produce a moderately skillful forecast if the area is not too restricted. That is, it appears that reasonable skill could be shown if forecasts are issued for specific areas roughly 30–40 km in radius. With the deployment of a national network of lightning ground-strike detectors, it seems reasonable to suggest that this type of forecasting deserves more research and development.

These experiments were purposely designed to make verification simple. Nevertheless, our experiences suggest that even when operational forecasters are included in the experimental design process, unforeseen complications and subtleties are almost certain to arise in operational practice. This suggests that more refined experiments than these (e.g., attempting to draw probability contours rather than assigning a single probability to an entire area) are likely to create difficulties for the experimenters. We advocate going ahead with such programs in the future, however; it is only by such mistakes that we will learn how best to accomplish the desired refinements.

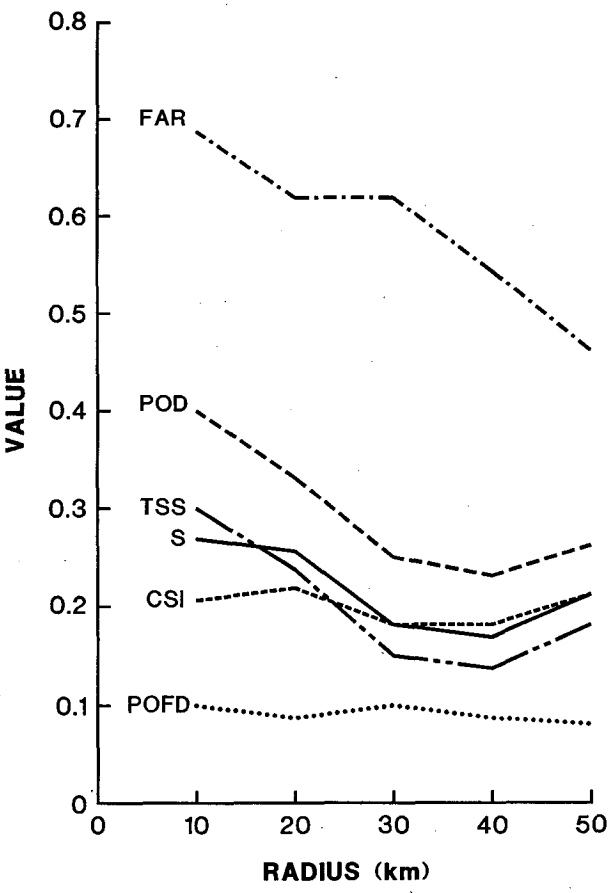

FIG. 3. As in Fig. 2, except for the Advance Outlook lightning forecasts.

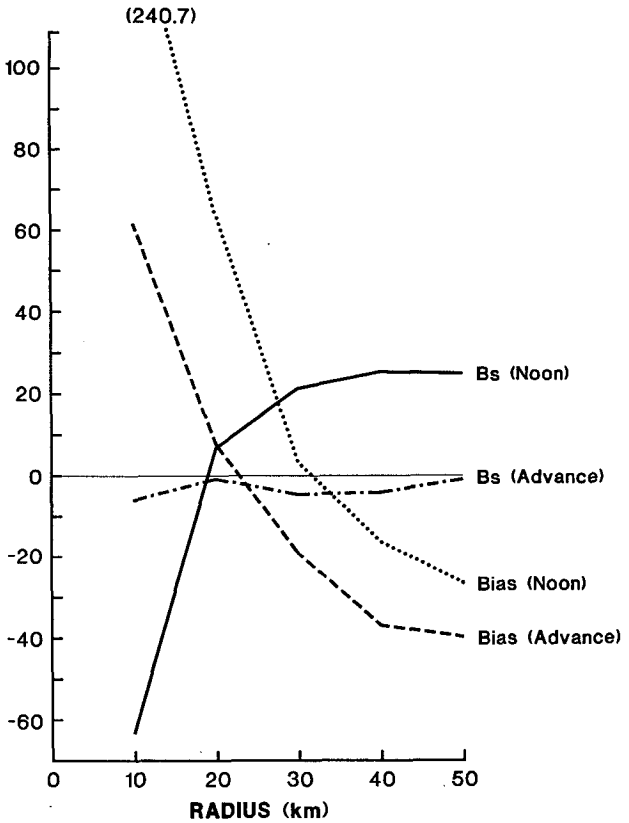

FIG. 4. Verification scores for both the Noon Outlook and Advance Outlook lightning probability forecasts as a function of the radius.

It is a continuing source of frustration for us that such things as defining a verifiable convective mode have been so difficult. What seems so clear during a discussion around a table before the experiment can become quite confusing after forecasting begins. Classification schemes that depend on qualitative characteristics of meteorological processes are not likely to provide a useful basis for forecast verification. On the one hand, such things as mesocyclones have turned out to be relatively easy to use in forecast verification because they are defined in terms of thresholds associated with measurable quantities (if a Doppler radar is available!). On the other hand, to date, most classification techniques for convective processes are rather subjective and do not have generally agreed-upon thresholds for those quantities one might measure in an operational experiment. For example, radar reflectivity or cloud-top blackbody temperature do not provide simple thresholds for deciding whether or not a convective cloud is thundering. Even CG lightning-strike detection may not do well at making such a distinction; it is well known that a considerable amount of lightning activity beyond CG strikes is present in convection, often well before the first ground strikes. Perhaps convective-scale vertical motion might serve

well, but it is not available. Despite the vexing aspects of these experiences, we believe that the difficulties will be worked out only by continuing such experiments.

Relative to our goals stated in the Introduction, we feel that the experimental forecasts have been successful, at least in part. There can be little question that, at least for some of us, the interaction between research and operational forecasting has been enjoyable and worthwhile. The most obvious evidence of that can be found in that we have done additional forecasting experiments in 1990 and 1991 (the results of which will be reported upon in the future). Regarding the research goal, the verification results presented herein are only the barest of beginnings. As noted in DF89, it is not until the reasons *behind* the verification results have been investigated and reported upon that the research can be said to have been successful. In view of the need to focus on trying to understand how the verification results came about, the type of forecasting experiments described in this paper will be suspended indefinitely after 1991. This should not be interpreted as indicating a failure, but merely as a desire to concentrate our resources on the purely meteorological aspects of our experiments. It is in this arena that we hope to make progress toward our basic and applied research goal.

Acknowledgments. The authors would like to thank all the members of the forecast teams from the Norman NWSFO and from NSSL for their contribution of time and effort in these experiments. We also want to acknowledge the encouragement and resource support of Dr. R. A. Maddox (NSSL) and Dr. K. C. Crawford (area manager, Norman NWSFO during the experiments). The senior author wishes to express his appreciation to Dr. Maddox and Mr. Tang Xin-zhang (Shanghai Meteorological Center) for making his extended visit to NSSL possible. Ms. Joan Kimpel's help with drafting the figures is appreciated, as is the constructive criticism of Dr. Ron Reap (TDL) and the anonymous reviewers.

REFERENCES

Barnes, S. L., 1985: Omega diagnostics as a supplement to LFM/MOS guidance in weakly forced convective situations. *Mon. Wea. Rev.*, **114,** 2121–2141.

Bothwell, P. D., 1988: Forecasting convection with the AFOS data analysis program. NOAA Tech. Memo. NWS SR-122, NTIS Accession No. PB89 145940/AS), 92 pp.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78,** 1–3.

Burgess, D. W., and L. R. Lemon, 1990: Severe thunderstorm detection by radar. *Radar in Meteorology,* D. Atlas, Ed., Amer. Meteor. Soc., 619–647.

Donaldson, R. J., R. M. Dyer, and M. J. Krauss, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints, *9th Conf. Severe Local Storms,* Norman, Oklahoma, Amer. Meteor. Soc., 321–326.

Doswell, C. A. III, and J. A. Flueck, 1989: Forecasting and verifying in a field research project: DOPLIGHT '87. *Wea. Forecasting,* **4,** 97–109.

——, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–585.

——, ——, and ——, 1991: Reply to comments on "On summary measures of skill in rare event forecasting based on contingency tables." *Wea. Forecasting,* **6,** 403.

Foster, M. P., 1988: Upper air analyses and quasi-geostrophic diagnostics for personal computers. Unpublished report, Scientific Services Division, National Weather Service—Southern Region Headquarters, 819 Taylor St., Fort Worth, TX 76102, 31 pp.

Hughes, L. A., 1980: Probability forecasting—Reasons, procedures, problems. NOAA Tech. Memo. NWS FCST 24, NTIS Accession No. PB80 164353, 84 pp.

Mach, D. M., D. R. MacGorman, and W. D. Rust, 1986: Site errors and detection efficiency in a magnetic direction-finder network for locating lightning strikes to ground. *J. Atmos. Oceanic Technol.,* **3,** 67–74.

Maddox, R. A., 1980: Mesoscale convective complexes. *Bull. Amer. Meteor. Soc.,* **61,** 1374–1387.

Murphy, A. H., 1991a: Probabilities, odds, and forecasts of rare events. *Wea. Forecasting,* **6,** 302–307.

——, 1991b: Comments on "On summary measures of skill in rare event forecasting based on contingency tables." *Wea. Forecasting,* **6,** 401–402.

Orlanski, I., 1975: A rational subdivision of scales for atmospheric processes. *Bull. Amer. Meteor. Soc.,* **56,** 527–530.

Reap, R. M., and D. S. Foster, 1979: Automated 12–36 hour probability forecasts of thunderstorms and severe local storms. *J. Appl. Meteor.,* **18,** 1304–1315.

——, ——, and S. J. Weiss, 1982: Development and evaluation of an automated convective outlook (AC) chart. Preprints, *12th Conf. Severe Local Storms,* San Antonio, Texas, Amer. Meteor. Soc., 110–115.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.,* **2,** 191–201.

Weisman, M. L., and J. B. Klemp, 1986: Characteristics of isolated convective storms. *Mesoscale Meteorology and Forecasting,* P. Ray, Ed., Amer. Meteor. Soc., 331–358.