Shootout–89, A Comparative Evaluation of Knowledge-based Systems That Forecast Severe Weather

W.R. Moninger[*], J. Bullas[**], B. de Lorenzis[***], E. Ellison[+], J. Flueck[++], J.C. McLeod[***], C. Lusk[*], P.D. Lampru[°], R.S. Phillips[°°], W.F. Roberts[*], R. Shaw[#], T.R. Stewart[##], J. Weaver[°°], K.C.Young[°],S.M. Zubrick[@]

Abstract

During the summer of 1989, the Forecast Systems Laboratory of the National Oceanic and Atmospheric Administration sponsored an evaluation of artificial-intelligence-based systems that forecast severe convective storms. The evaluation experiment, called Shootout-89, took place in Boulder, Colorado, and focused on storms over the northeastern Colorado foothills and plains.

Six systems participated in Shootout-89: three traditional expert systems, a hybrid system including a linear model augmented by a small expert system, an analogue-based system, and a system developed using methods from the cognitive science/judgment analysis tradition.

Each day of the exercise, the systems generated 2–9-h forecasts of the probabilities of occurrence of nonsignificant weather, significant weather, and severe weather in each of four regions in northeastern Colorado. A verification coordinator working at the Denver Weather Service Forecast Office gathered ground-truth data from a network of observers.

The systems were evaluated on several measures of forecast skill, on timeliness, on ease of learning, and on ease of use. They were generally easy to operate; however, they required substantially different levels of meteorological expertise on the part of their users, reflecting the various operational environments for which they had been designed. The systems varied in their statistical behavior, but on this difficult forecast problem, they generally showed a skill approximately equal to that of persistence forecasts and climatological forecasts.

1. Introduction

A comparative study of artificial intelligence (AI) systems that forecast severe weather was first proposed at the second workshop on Artificial Intelligence Research in Environmental Science (AIRIES) in 1987 (Moninger et al. 1987). At that time, it was noted that several prototype AI systems had been developed to produce these forecasts. Although the systems varied in the methodologies they employed and in operational environments they targeted, it was thought appropriate to bring them together for a comparative evaluation.

Ultimately, the developers of six different forecasting systems (not all of them AI) planned and executed an experiment called Shootout-89. In spite of the competitive-sounding name, the goals of the experiment were not to declare winners and losers; the systems are too immature for that to be useful. Rather, our goals were the following:

- To determine how best to compare diverse automated and semi-automated forecast methods.
- To determine the properties of the forecasts made using the different methods.
- To exercise the prototype systems and learn how to improve them.

Our desire was not to set up a "man versus machine" test. We believed this would set up such a competitive environment that it would distort the goals of the experiment. Therefore, we chose a forecast task for which, currently, no human forecasts are made: 2–9-h forecasts of severe and significant weather in four regions in and near the foothills of the Colorado Rocky Mountains. The task was thought relevant, however, because the National Weather Service is considering instituting forecasts of this type in the next several years.

This forecast task turned out to be more difficult than we anticipated; none of the systems produced particularly skillful forecasts. We believe we understand some of the particular reasons for the difficulties

[*] National Oceanographic Atmospheric Association (NOAA), Forecast Systems Laboratory, 325 Broadway, Boulder, CO 80303
[**] Atmospheric Environmental Service (AES), Arctic Weather Centre, Twin Atria Bldg., 4999 98th Ave., Edmonton, Alberta T6B 2X3, Canada
[***] Atmospheric Environmental Service, Forecast Research Division, 4905 Dufferin Street, Downsview, Ontario M3H 5T4, Canada
[+] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Campus Box 216, Boulder, CO 80309-0216
[++] University of Nevada at Las Vegas, Environmental Research Center, 4505 S. Maryland Parkway, Las Vegas, NV 89154
[°] Consultant's Choice, Inc., 8800 Rowell Road, Suite 130, Atlanta, GA 30350
[°°] NOAA/NESDIS/RAM Branch, Colorado State University, Foothills Campus, CIRA Bldg., Fort Collins, CO 80523
[#] Micro Forecasts, Inc., 319 SW Washington, Suite 909, Portland, OR 97204
[##] Center for Policy Research, Milne 300, State University of New York–Albany, 135 Western Avenue, Albany, NY 12222
[@] NOAA/National Weather Service, Office of Meteorology, 8060 13th Street, Silver Springs, MD 20910

©1991 American Meteorological Society

we encountered, and describe below what these reasons are and how we intend to address them in planned future exercises.

The small size of our sample (20 severe and 59 significant weather events) limits the statistical strength of the conclusions we can draw. Nonetheless, we believe a discussion of the issues we have addressed, and of the limited conclusions we can make, will be of use to others who may wish to plan comparative experiments of various human and automated forecasting methods.

2. Artificial intelligence

AI is a term coined in the late 1950s by John McCarthy, a computer science professor at Carnegie–Mellon University. A useful definition of AI is that by Raj Reddy, past president of the American Association for Artificial Intelligence. By his reckoning, AI is what you get when you combine problem-solving, computers, and heuristic knowledge. Heuristic knowledge is uncertain knowledge; it can vary from simple rules of thumb to sophisticated mathematical algorithms that are thought but not proven to converge. More traditional computer science generally studies the properties of well-defined algorithms; AI seeks to use ill-defined, uncertain, and approximate knowledge to solve problems. This might seem to be a step backward, but heuristic knowledge is what humans often use effectively to solve a great range of problems.

In Shootout-89, the particular AI technology that concerns us is that of expert systems. These are computer programs that consist of three distinct parts: a knowledge base, working memory, and an inference engine.

The knowledge base is a representation of the knowledge of one or more human experts about how to perform a task. The process of building the knowledge base, called knowledge engineering, involves acquiring the knowledge from the expert or experts, and developing a representation scheme that is complex enough to embody the salient aspects of the knowledge, but simple enough for the computer to generate solutions in a reasonable time. Representation schemes often consist of heuristic "if–then" rules, such as: "IF a strong capping inversion is present, THEN decrease the probability of significant weather." Knowledge representation may be a particularly difficult task when the knowledge is highly spatial in nature, as it is for weather forecasting.

Working memory is a database of current facts: those provided from the outside world via users, instruments, and other computer systems, and those inferred by the expert system. For example, a current fact might be "the observed dewpoint in Boulder is 63°F."

The inference engine is generally the only part of the expert system that actually does any computing. It compares the knowledge base to the facts in working memory, infers new facts, and draws conclusions. Often, the inference engine is designed to keep a record of what rules it used to draw conclusions, and it can report this information on demand, thereby "explaining" its conclusion to the human user.

Although some AI systems are designed to be totally objective and automated, some are designed to augment human skill. In these systems, the operator may provide subjective judgments as a part of the input. The final output depends on both the skill of the operator and the skill stored in the knowledge base.

To evaluate the performance of an expert system, one should consider the following issues.

- If subjective input is required, how sensitive is the system to that input? That is, how much of the resulting forecast depends on the skill of the operator and how much on the skill of the expert system?
- Is the mere use of the expert system (gathering the relevant data, making the requested decisions, if any) helpful for the operator in clarifying his or her decisions?
- If an explanation is provided by the system, is the explanation useful as a training or clarification tool for the user?
- How skillful is the ultimate forecast? If the forecast is less than perfect, is it because inappropriate subjective data were provided, incorrect objective data were provided, or the forecast model represented in the knowledge base is incorrect or insufficient?

3. Participating programs

Six systems participated in Shootout-89. Three were traditional expert systems, one was a hybrid system including a linear model augmented by a small expert system, and two were based on linear models. Three of the systems required varying amounts of subjective, user input, and the other three were (or could have been, if we had chosen to do it) totally automated. Table 1 describes the properties of the six systems.

The systems used data that are commonly available in operational weather service offices: numerical weather prediction (NWP) forecasts and analyses, soundings, and old surface observations. In addition, several systems used data from the PROFS mesonet, a network of 22 automated surface weather stations covering the experimental area.

TABLE 1. Summary of system properties

System	Architecture	Data Required	User Provides...	Platform
KASSPr[1]	Expert system	NWP analysis and forecasts	Subjective judgments of locations of meteorological features, drawn graphically	Linked DEC and Hewlett Packard workstations
Convex[2]	Expert system	Denver 1200 UTC sounding; mesonet data	Subjective forecast judgments (expected afternoon high temperature and mixed dewpoint, stability, moisture trends), provided via keyboard	PC
Willard[3]	Expert system	Denver 0000, 1200 UTC soundings; NWP analysis and forecasts	Subjective judgments of current meteorological conditions and trends, provided via keyboard	PC
GOPAD[4]	Multiple linear models produced from non-linear multiple-discriminant analysis of historical cases	1200 UTC soundings; NWP analysis and forecasts	Nothing—system is automated (previous day's verification is required for learning version)	Vaxstation
OCI[5]	Linear model augmented by a small expert system	Denver 1200 UTC sounding; NWP forecasts mesonet data	Objective answers to up to 30 questions provided via keyboard (could be automated)	PC
ALPS[6]	Linear model of 6 variables developed using "judgment analysis" methods	Denver 1200 UTC sounding; mesonet data	Objective answers to 6 questions in each region provided via keyboard (could be automated)	PC

[1] Developed by Carr McLeod, Bruno de Lorenzis, and John Bullas, at the Atmospheric Environment Service of Environment Canada, in cooperation with Digital Equipment Corporation.
[2] Developed by John Weaver and Roger Phillips at the NOAA/National Environmental Satellite, Data, and Information Service/RAMM branch.
[3] Developed by Steve Zubrick at Radian Corporation and at the NOAA/National Weather Service.
[4] Developed by Kenneth Young at the University of Arizona, in cooperation with Consultant's Choice, Inc.
[5] Developed by Robert Shaw, Thomas Corona, Denice Walker, and others at NOAA/Program for Regional Observing and Forecasting Services.
[6] Developed by Tom Stewart at the University Center for Policy Research, State University of New York at Albany, and Cynthia Lusk of the Center for Research on Judgment and Policy, University of Colorado.

a. KASSPr

In KASSPr, knowledge was elicited in a series of interviews and exchanges of documentation between the developer (de Lorenzis) and an expert in severe weather forecasting (Bullas). Demonstrations of the system were combined with additional knowledge acquisition sessions. At a fairly early stage, a prototype of the system was delivered to the expert for the running of historical test cases.

KASSPr was designed to be used in the environment typically found in an operational weather service office. KASSPr requires the meterologist operating the system to identify and analyze (draw) on the computer screen the forecasted positions of numerous meteorological features such as fronts and pressure, thermal, and vorticity troughs and ridges (de Lorenzis 1988). After this interaction, the system generates severe weather forecasts without further intervention.

KASSPr first evaluates the meteorological situation for necessary conditions. These are relatively few, but they must all be met for a given area or point to warrant further consideration. Next the situation is evaluated for sufficient conditions. Finally, a set of modifying conditions is applied to the probabilities and the severity factors.

b. Convex

For Convex (Weaver and Phillips 1987), Weaver, who has considerable severe-storm forecasting experience, provided the knowledge. Phillips, acting as the knowledge engineer, developed the rules using an expert-system building tool called EXSYS and provided the necessary linkages between the rules and the external processing model and database. Convex is designed to be used by moderately experienced meteorologists.

Convex first uses an automated analysis of the Denver morning sounding, combined with estimates of the expected afternoon temperature and dewpoint, to determine the relative instability of the host air mass and its likelihood of initiating convection later in the day over the region of interest. Later, it uses the most recent surface mesonet temperature and dewpoint measurements and a linear, time-dependent, boundary-layer mixing function to derive updated values for these same two parameters for each of the subregions of interest.

The meteorologist operating Convex may override Convex's estimates of the low-level, mixed afternoon dewpoint to provide his or her own. The operator must also provide reasonably knowledgeable information about synoptic-scale conditions.

On request, Convex will display a backward sequence of the rules that were evaluated to be true, and thereby explain its reasoning.

c. Willard

The knowledge base for Willard is a structured hierarchy of 30 rules. Most of the rules were developed using the inductive generalization feature of RuleMaster, an expert system shell. Examples of forecaster decision-making were fed into RuleMaster, and the decision rules induced were examined by the developer (meteorologist) for suitability and correctness. The rules were subsequently modified by hand. For Shootout-89, additional rules were added that pertain to shorter-range, severe-thunderstorm forecasting as practiced by forecasters at the Denver NWS Forecast Office. Like Convex, Willard can provide explanations of its reasoning.

Willard was designed for use by novice meteorologists. The user must provide subjective information about current synoptic and mesoscale features and interpretation of numerical forecast guidance.

d. GOPAD

GOPAD refers to the software used to extract patterns from historical data to create a forecast model from these patterns. The pattern extraction is based on multiple discriminant analysis (MDA) techniques developed by Miller (1962) and is extended to provide nonlinear discrimination. In addition, the GOPAD software creates combinations of simple predictor variables which are termed indices; the indices are used in the nonlinear MDA analysis.

Forecasts are produced by finding a set of days in the historical dataset that are analogous to the forecast day. Probability forecasts for a region are based on the historical frequency of occurrence of each forecasted weather category in the set of analogues for each region. Six subforecast models were used to produce three subforecasts for each region. The issued forecast is the median of the three subforecasts.

More than 1000 potential predictor variables drawn from observed rawinsonde and mesonet parameters were analyzed with the GOPAD software. A total of 268 days from the summers of 1983, 1985, and 1987 were used to create the forecast models used in Shootout-89.

Two versions of GOPAD were run in Shootout-89. The "learning" version received verification information for each previous forecast day and added that to the historical database along with the previous set of predictor variables. The "static" version maintained the original historical database during the course of the experiment. Both versions were designed to be operated by nonmeteorologists.

e. OCI

Many Boulder meteorologists provided the knowledge used by OCI. First, a list of potential predictors was compiled; however, the archived data were insufficient and the number of potential predictors too large for effective regression equations to be generated. Instead, predictors were subjectively weighted and a linear model was built. A small expert system was added to identify and account for relationships among variables that might inhibit convection. OCI was not designed to generate forecasts for the mountain region.

OCI was designed to provide automated forecasts. No meteorological expertise is required of the operator. Because the version used in Shootout-89 did not have automatic data ingest, a small amount of knowledge about the meteorological infrastructure was required in order to identify the needed products.

f. ALPS

ALPS is based on psychological research on judgment and decision making that has repeatedly shown that, under certain conditions, simple algebraic models can capture the skill of expert judgment, and often outperform the expert. (See, e.g., Dawes et al. 1989.)

The development of ALPS began with the identification of a set of precursors, or "cues," thought to be important predictors of convection. Potential cues were gathered using the traditional cognitive science techniques of structured interviews and a structured meeting of several meteorologists, facilitated by a cognitive scientist. The cues selected were positive buoyancy, wind shear, dewpoint, surface wind direction, and surface temperature. Measures for each cue were then developed, and weights assigned so that each cue would have a roughly equal impact on the forecast. The forecasts issued by ALPS were weighted sums of the cues, calibrated by region using the data from the summers of 1983, 1985, and 1987 (see Table 6).

ALPS was designed to be operated by nonmeteorologists. However, in use, it was noted that some meteorological skill was necessary to interpolate when required data were missing.

4. Experiment design and operations

The location for Shootout-89 was the high plains and foothills of northeastern Colorado. This region was chosen because the Forecast Systems Laboratory (FSL) Program for Regional Observing and Forecasting Services (PROFS) has been conducting forecast exercises there for several years, and therefore an extensive mesoscale retrospective dataset exists, as do extensive verification data. In addition, PROFS maintains a network of automated surface observing platforms (the PROFS mesonet), which provides valuable data for regional weather forecasting.

Shootout-89 ran from 15 May until 17 August 1989. Each weekday of the exercise, each system (with the exception noted below) generated mutually exclusive and exhaustive probabilities that the most severe weather in each of four regions would be in one of three categories:

Category 0: Nonsignificant weather (nil), defined as the absence of category 1 or 2 weather.

Category 1: Significant weather (sig), defined as a storm that has any of the following: hail with a diameter between 0.25 and 0.74 inches, surface winds between 35 and 49 kt, 2 inches per h or greater rainfall rate, or a funnel cloud aloft.

Category 2: Severe weather (svr), defined as a storm observed to have at least one of the following: hail with diameter equal to or greater than 0.75 inches, surface winds of 50 kt or greater, or a tornado.

As shown in Fig. 1, the forecast area was divided into four climatologically distinct forecast regions based on work by Weaver and his colleagues (Weaver and Phillips 1987; Weaver et al. 1987; Klitch et al. 1985). The systems produced forecasts for each region. The exception was OCI, which did not generate forecasts for region 1. Willard generated the same forecast for each of the four regions.

Forecasts from all systems except OCI were finished by approximately 1115 MDT (1715 UTC). Because of the amount of manual input required, OCI was run in the early afternoon using morning data. The valid time for the forecasts was 1300–2000 MDT (1900–0200 UTC).

The systems were operated by either the chief meteorologist (CM) or the backup meteorologist (M2). On a few days (discussed below), both the CM and M2 operated the systems independently. Both the CM and M2 are moderately experienced research meteorologists. The CM operated the systems on approximately 60 days; M2 operated the systems on approximately 9 days. M2 never ran OCI, because it was run in the afternoon when the CM was always on duty.

The activities of the CM were monitored on six randomly chosen days of Shootout-89. He was aware of being monitored; we cannot assess the extent to which this awareness might have affected his behavior. He was well-acquainted with the observer, however, and visitors and informal observers were often in the Shootout forecast room. These facts suggest that the effect of the observer on the CM's behavior would be slight.

The CM started work at about 0900 MDT (1500 UTC) when initial data to be used by the programs first became available. The CM spent the first hour on system startup and data acquisition, including logging on to the computers, setting up data links, obtaining sounding, mesonet, and the previous day's verification data, and looking at these data as well as maps and satellite images. Between 1005 and 1030 MDT (1605 and 1630 UTC), the NWP and final mesonet data necessary to run the programs became available, and the CM began providing the necessary objective and subjective input to the programs.

The CM was free to run the programs in any order, and in fact the order in which the programs were run was different for each of the six days monitored. The mean times spent for different types of activities during the six days are presented in Table 2. It is important to note that the time shown for examining the data may be inflated from what is actually necessary to run the programs because on some days the CM had to wait for data and may have spent time looking at other data while waiting. Table 2 does not include time spent logging on and transferring data between machines, dealing with problems, completing paperwork, and performing other organizational activities.

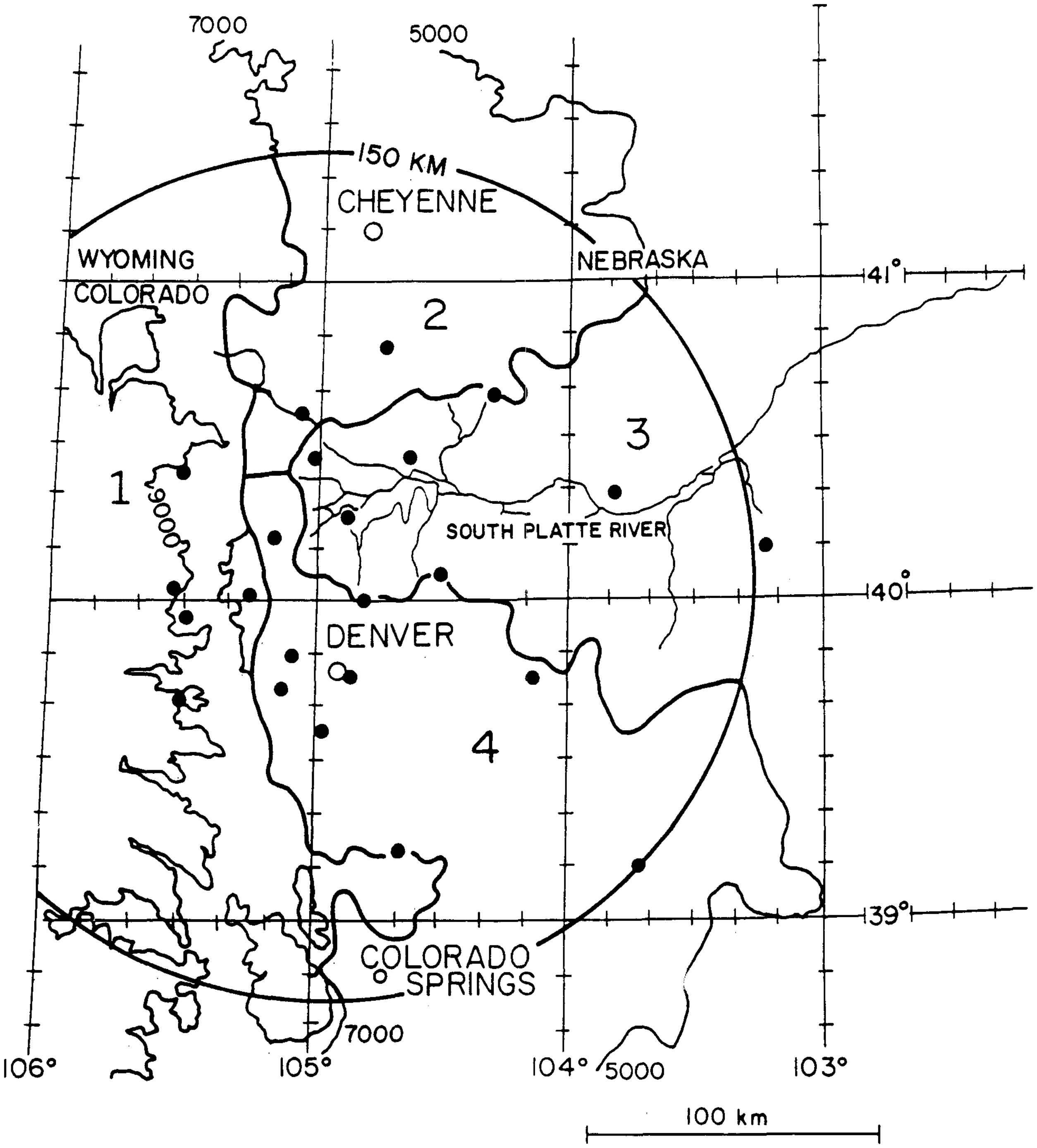

FIG. 1. Four regions of Shootout-89.

At approximately 1125 MDT (1725 UTC), M2 presented forecast results from the AI systems at the FSL daily weather briefing. At this time, verification data from the previous forecast day were also presented.

The backup meteorologist (M2) reported that her activities did not differ materially from those of the CM.

5. Evaluation of utility

In this section, we discuss several kinds of nonmeteorological evaluation, of both system behavior and our operational procedures, that do not depend on verification data; evaluations that do depend on verification data are discussed in the following section.

TABLE 2. Operator time (in minutes) spent on Shootout activities

Activity	Mean	SD
Obtaining sounding, mesonet, and verification data	16.42	6.80
Looking at data	17.33	5.58
Running programs (not including OCI)	29.79	3.21

a. Sensitivity to operator input

To what extent do different operators of the same system, using the same data, produce similar forecasts? GOPAD, ALPS, and OCI require purely objective input data, and therefore have no sensitivity to operator input. KASSPr, Convex, and Willard, on the other hand, require subjective interpretation of data and/or subjective determination of input values; for these three systems, differences in the subjective processing of information may result in different output forecasts.

Two strategies were used to assess the sensitivity of systems to operator input. The first was to compare the forecasts generated when the three systems were run by the CM to those generated when the systems were run by M2. The second strategy was to have M2 run the three systems retrospectively, using data from selected days. These days were selected because some significant or severe weather had occurred and they had been ones on which the systems had produced differing forecasts. When running the systems retrospectively, M2 had all the necessary objective data on computer files as well as hard copies of other data available to the CM when he ran the systems (including PROFS mesonet data, sounding analysis package output, satellite images, and traditional weather maps). The forecasts analyzed include those from the two days when M2 ran the three systems after the CM had completed his day's activities, and the two days for which M2 ran the systems retrospectively. Across the four regions, this yielded 16 forecasts for each forecast type (nil, significant, severe) for each of the three systems.

We first computed the number of forecasts for which pairs of forecasters "agreed." Two levels of agreement were considered: within a probability of 0.05 and within a probability of 0.20. The stringent criterion was chosen to represent a level of disagreement that we considered would not be meaningful to users. The lenient criterion was intended to represent a level of disagreement that was more meaningful. The number of times (out of the 16 possible) the CM and M2 were within the two agreement levels are reported in the first two columns of Table 3. We next calculated the mean difference in forecasts, the mean absolute difference in forecasts, and the maximum absolute difference in forecasts. These are reported in the third, fourth, and fifth columns of Table 3, respectively.

Table 3 indicates the extent to which, for the days in our sample, the systems were sensitive to operator input. It must be noted, first, that the amount of data we have limits the ability to generalize our results, and second, that the selection criteria for the days in this sample are likely to maximize the sensitivity to operator input. Our limited data suggest that all three systems can be quite sensitive to operator input. Our primary goals in presenting these data are to suggest that sensitivity to operator input may be a concern in some operational settings, that systems vary in their sensitivity to operator input, and that if this is a concern then the degree of sensitivity should be addressed. Our research provides an example of how such effects could be addressed, but our sample does not allow any strong conclusions.

b. Participant evaluation of systems

Near the end of the field season, the two operators were asked the extent to which they were satisfied with various aspects of each system, how well they understood the system, how useful they felt the system was in helping them understand the weather, and if they felt the system was best suited for operational or training environments.

Each operator filled out a detailed questionnaire, the results of which are summarized in Table 4. A weighted average of the responses of both operators was used to generate the results shown. Items for which their responses diverged substantially are noted below. The operators were instructed to "try to answer each question *without regard to skill* of the system."

The difference between the systems requiring subjective input (KASSPr, Convex, and Willard) and the others is evident in the first entry of Table 4. Of the objective systems, only ALPS required any meteorological skill to operate, and this was only to interpolate when data were missing.

Both operators found all systems to be quick and easy to use. KASSPr was thought to be somewhat more difficult to use than the other systems because of the need to provide extensive graphical input.

TABLE 3. Sensitivity to operator input as analyzed by comparing forecasts of CM and M2 (n=16)

Forecast by system	Within a probability of 0.05[1]	0.20[2]	Mean diff[3]	Mean adiff[4]	Max adiff[5]
Nil					
KASSPr	9	10	2	20	68
Willard	12	12	-10	10	40
Convex	4	12	0	16	60
Sig					
KASSPr	9	10	8	18	64
Willard	12	12	8	8	30
Convex	4	11	4	17	50
Svr					
KASSPr	10	12	-10	12	75
Willard	12	16	3	3	10
Convex	13	15	-4	6	60

[1]Number of differences less than or equal to 0.05.
[2]Number of differences less than or equal to 0.20.
[3]The mean of the differences in forecasts.
[4]The mean of the absolute differences in forecasts.
[5]The maximum of the absolute differences in forecasts.

Question 4 was asked to assess the extent to which the operators understood the knowledge underlying each system's forecasts. It should be noted that there was no attempt to formally train the operators about the knowledge within each system, so their responses represent insights they picked up while running each system and talking informally with system developers. On this question, operator opinions diverged. The chief meteorologist indicated best understanding of KASSPr and Willard, followed by Convex, with less understanding of the other systems. The backup meteorologist indicated moderate understanding of ALPS, Convex, and KASSPr, and low understanding of GOPAD and Willard.

Question 5 addresses how well the systems aid an operator in understanding synoptic weather. KASSPr was thought very useful, indicating that the time necessary to enter the required graphical input data provided a payoff in understanding. Willard was also considered useful, Convex less so, and the three objective systems were not considered useful in providing synoptic understanding.

Question 6 addresses how well the systems aid an operator in understanding mesoscale weather. Convex and Willard were thought most useful, followed by KASSPr, OCI, and ALPS. Although OCI and ALPS required only objective input, the operators apparently believed that simply assembling the needed input data provided some mesoscale understanding. GOPAD, being entirely automated, was, of course, not considered useful in providing mesoscale understanding.

Questions 7 and 8 address how useful the systems might be in an operational weather service environment. As forecast aids, all systems were thought at least moderately useful, with KASSPr, Convex, and OCI considered more potentially useful than the others. As training aids, not surprisingly, the systems requiring subjective input were considered useful, and the other systems were considered only marginally useful.

Finally, question 9 assesses overall satisfaction. The operators were most satisfied with KASSPr and Convex, were dissatisfied with ALPS, and were neutral about the others.

6. Evaluation of statistical skill

a. Verification data

The collection of verification data was a crucial aspect of Shootout-89. A full-time verification coordinator (VC) gathered and documented verification data for the exercise. During times of expected significant or severe weather, the VC was stationed at the Denver

Table 4. Operator feedback

Questions	KASSPr	Convex	Willard	GOPAD	OCI	ALPS
1. "Check the operator skills necessary to run the system."						
Interpret sounding data	X	X	X			
Identify current mesoscale features	X	X	X			(X)
Identify current synoptic features	X	X	X			
Forecast mesoscale features	X	X	X			
Forecast synoptic features	X	X	X			
2. "Rate the overall ease of use of the system."	Moderately easy	Quite easy	Quite easy	Quite easy	Very easy	Very easy
3. "Rate the general timeliness/ fastspeed of running the system."	Moderately fast	Quite fast	Quite fast	Quite fast	Quite fast	Quite fast
4. "Assume that you and the system each generated a forecast independently and that the forecasts were discrepant. How readily could you explain the basis of the discrepancy?"	Well	Well	Well	Poorly	Poorly	Poorly
"To what extent is the system useful as an aid in understanding a given day's:						
5. Synoptic weather?	Very useful	Marginally useful	Quite useful	Not useful	Not useful	Not useful
6. Mesoscale weather?"	Moderately useful	Quite useful	Moderately useful	Not useful	Moderately useful	Moderately useful
"Assume you had the option of installing the system in the local WSFO:						
7. How useful would the system be to operational meteorologists generating forecasts?	Quite useful	Quite useful	Moderately useful	Less useful	Moderately useful	Less useful
8. How useful would the system be in training new operational meteorologists?	Quite useful	Quite useful	Moderately useful	Less useful	Less useful	Less useful
9. Rate your overall satisfaction with the system."	Quite satisfied	Quite satisfied	Neutral	Neutral	Neutral	Somewhat dissatisfied

WSFO, where there was access to real-time radar data. When radar or other data suggested possible significant or severe weather, the VC called cooperating observers in potentially affected regions. The VC also received reports phoned in to the WSFO and, on days following possible weather events, made follow-up phone calls.

The following sources provided verification data: 1) a volunteer spotter network and a paid cooperative observer network sponsored by the NWS; 2) police and fire stations, county emergency preparedness staffs, and highway road crews; 3) a network of amateur radio operators; 4) weather service offices in Colorado Springs and Cheyenne, Wyoming; 5) 22

TABLE 5. Operational days

System	Starting date	Operational days
KASSPr	15 May	60
Convex	15 May	57
Willard	30 May	54
OCI	30 May	50
GOPAD (learning)	15 May	53 (+10)*
GOPAD (static)	22 May	50 (+11)*
ALPS	7 June	49 (+7)*

*Additional days rerun after the completion of the field season are indicated parenthetically.

automated mesonet stations operated by PROFS that provided data on maximum wind gusts and rainfall rate; 6) daily weather observations recorded by approximately 30 specially recruited weather observers who mailed in observations monthly; and 7) occasional volunteer chase teams of research meteorologists.

We define a case as a weather classification for a single region or a given day, e.g., severe for region 1 on 20 June. The severity of a case is determined by the most severe weather that is reported in the region on each day. The distinction between an individual weather event (i.e., a report) and a case should be kept in mind in the discussions to follow.

At the end of the field season, another meteorologist analyzed all the event documentation and radar data that had been gathered and used by the VC. As a result of this analysis, the initial verification determinations of the VC were changed in only 7 cases out of all 276 cases (4 regions multiplied by 69 days). This gives us considerable confidence in the coding and evaluation of spotter reports. Nevertheless, significant and severe weather are rare events; it is likely that several storms were missed because there were no spotters in the area. Similarly, the distinction between significant and severe should be viewed with caution, because spotters may not have been in a position to observe the most severe portion of a storm. For this reason we grouped significant and severe weather together and called such cases "non-nil."

Several cases were verified as non-nil because of strong winds that did not appear to be associated with convective events. Although such events passed our criteria for significant or severe weather, we were concerned that the inclusion of such cases might decrease the statistical skill of systems designed to predict convective weather. To assess the extent of this possible effect, a specialist in Colorado windstorms (J. Brown of FSL) evaluated sounding, radar, mesonet, and satellite data associated with all wind events that affected the classification of a case, and declared each to be either convective or nonconvective. Removing nonconvective events would have changed the verification classification in 14 cases out of the 276 cases. Skill scores with the nonconvective cases removed were not significantly different than the results we present here, which include the nonconvective cases.

b. Sample size and climatology

Table 5 shows starting dates and the number of days that each system successfully generated forecasts during Shootout-89. Forecasts were not made when a program failed to run because of software problems or when required input data were not available. Two of the objective systems (GOPAD and ALPS) were rerun after the field season for some of the days for which they had been unable to generate forecasts. These additional days are included in the final sample.

Common days are those for which all systems generated forecasts. For region 1, in which OCI did not participate, there were 48 common days. For regions 2–4, there were 45 common days each. This yields a total sample of 183 cases. Most of the evaluation to follow applies to this sample of common days. Because results from the two versions of GOPAD were very similar, we present results from only the learning version.

Table 6 shows climatology for the Shootout-89 common days, and for the summers of 1983, 1985, and 1987. On those three summers, PROFS conducted extensive forecast-verification exercises using an extensive network of radio-directed chase teams.

The frequency of severe cases for Shootout-89 is similar to the frequencies observed in the three previous exercises. However, there is a substantial difference in the frequencies of significant cases. Differences between verification procedures in Shootout-89 and in the three PROFS exercises can explain some, but not all, of this difference. The PROFS exercises did not send chase teams into the mountainous portions of region 1; this could account for some of the frequency differences in that region. However, it cannot explain the substantial differences in the other regions. Nonconvective cases cannot account for the difference; their removal changes the base rate for significant cases in the common days sample only slightly—to 0.35, 0.33, 0.13, and 0.38 for the four regions, respectively.

We conclude that either significant weather was underreported in the PROFS exercises or we

TABLE 6. Observed frequencies (climatology) of each weather category for Shootout-89 common days and for summers of 1983, 1985, and 1987

	Region			
Weather Category	1	2	3	4
Nil				
Shootout-89	0.54	0.62	0.67	0.44
3 summers	0.90	0.88	0.85	0.72
Sig				
Shootout-89	0.38	0.33	0.20	0.38
3 summers	0.06	0.07	0.06	0.09
Svr				
Shootout-89	0.08	0.04	0.13	0.18
3 summers	0.04	0.05	0.09	0.19

overreported significant weather (possibly by asking leading questions). It is also possible that 1989 was a more active summer than the earlier years, and that severe events were underreported in 1989, but significant events were not. We believe that it is most likely that significant weather was underreported during the PROFS exercises; interviews with some of the principals of those exercises suggest that there was far more emphasis on reporting severe events than on reporting significant events.

It should be noted that if significant weather events were missed during the RT exercises, this would primarily effect the GOPAD system, which depends on the properties of known past weather events to make its predictions. To a lesser extent, ALPS, which was calibrated using past statistical data, would be affected. None of the other systems depend on the historical data from the RT exercises.

In some of the analyses reported here, we refer to climatology forecasts. These are forecasts in which the predicted probability of occurrence of each weather category in each region is taken as the observed past frequency of that category for that region. For our climatology forecasts, we used the 3-summer frequencies shown in Table 6.

Considerable regional variability is evident in the data shown in Table 6. There are several reasons for this variability: the regions are climatologically distinct (Weaver et al. 1987); the regions have different population densities, and hence different densities of potential weather reporters; and the limited sample size in each region results in relatively large statistical fluctuations. Because of the last reason, we aggregated the skill scores over regions.

c. Persistence forecasts

Persistence forecasts were generated by forecasting that, in each region, each day's weather would be the same as the previous day's weather. We were able to generate persistence forecasts for 45 days of the experiment for which we have verification information from the previous day. These days are not entirely the same as the common days.

Although the persistence forecasts are for a different sample of days than the common days, we include them in our analyses because the frequency of occurrence of non-nil cases is very similar in the two samples, as Table 7 shows.

d. Forecast skill

We divide forecast skill into two components: resolution and bias. Resolution measures the extent to which forecasted probabilities for weather categories are consistently higher (or at least different) when weather in that category occurs than when it does not. Bias measures the extent to which the average forecast for a weather category matches the observed relative frequency of that category.

1) Resolution

To assess the ability of the systems to resolve nil, significant, and severe weather, we use correlation measurements and signal detection theory (SDT). Table 8 shows the correlations between the forecasts and the observed weather. The two correlations by region shown were calculated using the entire common-days sample of 183 cases (135 for OCI).

The correlations by region are small and positive, but not generally statistically significant. If we consider reduction of variance, which is the square of the correlation, no system reduces the variance by more than 7%, telling us that the systems have ample room for improvement. Not all 183 cases are independent, because some of the relevant weather patterns extended over several regions. If we assume a smaller number of degrees of freedom, for example 100, only

TABLE 7. Frequency of non-nil cases

Dataset	**Region**			
	1	**2**	**3**	**4**
Common days	0.44	0.37	0.33	0.56
Persistence days	0.44	0.40	0.36	0.53

Table 8. Correlations between forecasts and observed cases

	Correlations		
System	By region[1]		Entire area[2]
	Non-nil	Severe	Severe
KASSPr	0.24	0.11	0.28
Convex	0.26	0.17	0.50
Willard	0.02	0.00	0.00
OCI*	0.08	0.12	0.23
GOPAD	0.15	0.21	0.18
ALPS	0.07	0.16	-0.04
Climatology	0	0	0
Persistence**	0.17	0.22	0.12

*OCI did not forecast for region 1

**Different set of days; see section 5.3

[1] Correlations for the common days sample (n=183 generally; n=135 for OCI).

[2] Correlations between each system's most severe forecast for each of the 45 common days and the most severe weather observed in any of the four regions (three regions for OCI).

correlations greater than 0.20 are significant at the p<0.05 level, two-tailed.

The first two columns of Table 8 suggest that, in our weather sample, KASSPr and Convex resolved non-nil weather, while GOPAD and Persistence resolved severe weather. However, our data are too limited to be more than suggestive.

We investigated the possibility that the poor showing of the systems at resolving severe weather might be due to the small size of the regions. To estimate how well the systems would have done had they generated forecasts for a larger region, we compared each system's highest probability severe forecast for each day with the most severe weather that occurred in any of the four regions. Those results, given in the third column of Table 8, are encouraging. Both Convex and KASSPr show modestly significant (n=45, p<0.07, two-tailed) abilities to resolve severe weather in the larger area. This suggests that the systems embody some understanding of the overall synoptic situation, but they do not yet have sufficient spatial precision to focus their forecasts well on the appropriate region.

Another way to study resolution is to use signal detection theory (Swets 1988; Mason 1982). In this approach, plots are made of each system's probability of detection versus probability of false detection. Using general assumptions about the statistical properties of the forecasts, a curve called the receiver (or relative) operating characteristic (ROC) can be fit to the data. Without going into further detail, it will suffice for our purposes to note that, if an ROC curve can be fit to the data, the general statistical assumptions of signal detection theory are satisfied, and the area under the best-fitting ROC curve is a measure of resolving power (Dorfman and Alf 1969). An area of 1.0 indicates perfect resolving power; an area of 0.5 indicates forecasts that are no better than chance.

Table 9 presents the area under each system's ROC curve. The areas were calculated using a program written by L. Harvey of the University of Colorado that applies the maximum likelihood fitting method of Dorfman and Alf (1969). Standard errors based on the fitting method are indicated.

The ROC results corroborate the correlation analysis. Within each forecast region, the systems show a slight (but probably not statistically significant) ability to resolve nil weather events. But when we compare weather over the entire experimental area, two of the systems, KASSPr and Convex, do show encouraging resolving power. Note that the rank order of the systems is different between the ROC analysis and the correlation analysis, even though the two analyses measure similar properties of the forecasts. This is a reflection of our limited statistical sample, emphasizing the need to consider multiple measures and to avoid statistically unwarranted conclusions. We do not show an analysis of nonsevere versus severe forecasts by region because the SDT results were too unstable to be useful.

2) Bias

The overall skill of a forecast model will be adversely affected if the model is biased; i.e., it produces forecasts that are too low (underforecasting) or too high (overforecasting). The bias is defined as the average forecasted probability for a given weather category divided by the observed frequency of that category. A bias less than (greater than) unity represents underforecasting (overforecasting). Table 10 shows the bias for each system for both non-nil and severe-weather forecasts. With one exception, all the models underforecast. The least-biased forecasts were produced by GOPAD for severe weather; the most-biased forecasts were produced by ALPS for non-nil weather. Some of the apparent underforecasting of non-nil weather may be because the observed frequency of significant weather was much higher in 1989 than in the three earlier summers used to calibrate GOPAD and ALPS (see section 5b). If significant weather were overreported in 1989, this

TABLE 9. Signal detection theory analysis summary

	Area under ROC curve	
System	By region[1] Nil vs. non-nil	Entire area[2] Non-svr vs. svr
KASSPr	0.74+/-0.15	0.81+/-0.24
Convex	0.65+/-0.05	0.74+/-0.12
Willard	0.53+/-0.05	**
OCI*	0.58+/-0.05	0.63+/-0.10
GOPAD	0.56+/-0.05	0.61+/-0.10
ALPS	0.55+/-0.05	0.55+/-0.17

*OCI did not forecast for Region 1.
**Fitting algorithm did not converge.
[1] For common days sample (n=183 generally; n=139 for OCI).
[2] For each system's most severe forecast for each of the 45 common days verified against the most severe weather observed in any region.

would decrease the bias of all systems for non-nil weather.

7. Overall measures of system performance

Table 11 presents a synthesis of the foregoing results. Column 1 indicates that KASSPr and Convex generally have the best resolving power, consistent with that of persistence. No system unequivocably beats persistence. The other systems do not exhibit statistically significant resolving power in our data sample; however, there is a suggestion that GOPAD and OCI do somewhat better than do Willard and ALPS.

Column 2 indicates that all the systems, with the exception of GOPAD and Convex, are substantially biased (they generally underforecast). Also, both persistence and 3-year climatology underforecast with respect to our sample climatology.

Combining the effects of resolution and bias leads to traditional skill scores that are generally low with respect to persistence. For instance, critical success indices (CSI[1]) (Donaldson et al. 1975) for forecasting non-nil weather range from 0.31 (Convex) to 0.09 (ALPS), with persistence scoring 0.33. For discriminating severe weather, CSIs range from 0.11 (OCI) to 0.02 (Willard), with persistence scoring 0.17.[1]

Column 3 of Table 11 indicates that, regardless of the skill of resulting system forecasts, running the three systems that require subjective input can be a useful exercise; those systems tend to lead the meteorologist/operators through an organized briefing on the meteorological situation (see section 4). Column 4 generally indicates the converse; OCI, GOPAD, and ALPS could be far more easily automated than the systems for which subjective input is critical. Convex and Willard have a partially automated mode, but were not specifically designed to be automated tools.

TABLE 10. System bias for non-nil and severe-weather forecasts

System	Non-nil	Severe
KASSPr	0.28	0.34
Convex	0.67	0.57
Willard	0.30	0.27
OCI*	0.71	1.51
GOPAD	0.38	0.81
ALPS	0.23	0.54
Climatology	0.38	0.85
Persistence	0.65	0.72

*OCI did not forecast for Region 1.

8. Discussion

The generally low skill of the systems can be attributed to two factors: the difficulty of the forecast task, and limitations in the systems themselves.

a. Difficulty of the forecast task

Forecasting severe-weather events with a 2–9-h lead time for predefined regions as small as those used by Shootout-89 substantially stretches the state of the art. No human forecaster has attempted this task operationally for the topographically diverse northeastern Colorado region. In retrospect, it is quite clear that, in our enthusiasm, we violated a cardinal rule of system development: start simple.

Afternoon forecasts have been made for similar

[1] CSI is generated from a contingency table. Traditionally, a series of thresholds is needed to convert probabilistic forecasts to the categories (nil, sig, svr) in the table. An alternate formulation that we have employed is to enter into each cell of the contingency table the sum of the forecasted probabilities for each of the observed cases. Using this formulation, no thresholds are required.

TABLE 11. Overall measures of system performance

System	Resolving Power	Bias	Suitability as a Briefing Tool	Suitability as a Stand-Alone Tool
KASSPr	1	3	1	3
Convex	1	2	1	2
Willard	3	3	1	2
OCI	2	3	3	1
GOPAD	2	2	3	1
ALPS	3	3	3	1
Persistence	1	2	—	—
Climatology	3	2	—	—

1=Relatively good
3=Relatively poor

spatial regions in northeastern Colorado (Heideman 1989). The skill scores reported for these forecasts are better than those of the Shootout-89 systems; however, we believe that this is due to a difference in forecast issue times. The human forecasts were made at 12:30 P.M. local time; Shootout-89 forecasts were based on data from 10:30 A.M. and earlier. McGinley et al. (1991) suggest that the boundary layer, which determines the surface data used by all of our systems, is much better coupled with the overall air mass after noon local time. Thus, the forecasts issued at 12:30 reported by Heideman may have had the advantage of being based on more relevant meteorological data.

An additional problem with the forecast task is that the quality of the verification data for significant events varied between the training sample used by several of the systems and the Shootout-89 verification data. Thus, the Shootout-89 forecast task was, in effect, different from what these systems had been trained for, at least for significant weather.

b. Limitations in the systems

KASSPr. After the end of the experiment, errors were found in the algorithms that converted spatial graphical input to symbolic data for use in the rules that established necessary conditions. The effect of these errors, and of the relatively poor temporal resolution (one NWP input at 0000 UTC only) was to cause KASSPr to forecast severe probabilities of 0.0 for several of the cases that verified as severe. Post-facto analysis of the behavior of individual rules corroborates this. The algorithms are being corrected, and the temporal resolution is being improved.

Also, KASSPr's dependence on NWP models that do not yet resolve mesoscale conditions well limited KASSPr's ability to forecast on the mesoscale. Indeed, KASSPr performs substantially better (Bullas et al. 1990) when forecasting for larger regions. Additional mesoscale rules are being developed.

Convex. The primary variable Convex was designed to forecast is buoyant potential energy. For Shootout-89, this variable was converted to significant and severe-weather probabilities discontinuously. This greatly amplified Convex's sensitivity to input data. For example, in several cases a change of 0.5°C in the input dewpoint forecast would have changed Convex's forecasted probabilities by 60%.

The effect of this artificial sensitivity was to depress Convex's measured accuracy, particularly for severe weather of which there were only a few cases. For example, in forecasts of severe, CSI values of approximately 0.1 seem to indicate poor accuracy. However, when significant and severe are taken together, CSI values climb to 0.31, which is more typical of similar forecasts in eastern Colorado (McGinley et al. 1991).

Post-facto tests of probability forecasts using desensitized thresholds suggest much-improved accuracy. More important, these new, graduated probabilities, based on several graduated thresholds, are probably a much more realistic representation of what can be known in advance concerning thunderstorm severity.

Sensitivity to input dewpoint forecasts still remains a factor, however; this is an important variable that is often estimated unreliably by the human operator. The new version of Convex will require the operator to use a decision-tree approach to assist in determining an afternoon mixed-dewpoint value.

Willard. Willard's performance was due, we believe, to the system being asked to forecast on the mesoscale using a synoptic-scale knowledge base. Willard was designed for 0–24-h forecasts of severe weather in larger regions than those used by Shootout-89. Indeed, Willard exhibited better performance (Zubrick and Riese 1985) on these larger temporal and spatial scales than it did in Shootout-89. Although some Colorado mesoscale knowledge was added to Willard, it was clear to the developer at the outset that limited resources would not allow the level of effort required to develop a sufficient mesoscale knowledge base. Willard's performance in Shootout-89 strongly suggests that its synoptic-scale knowledge was insufficient for adequate mesoscale forecasting.

OCI. No detailed study has been done of OCI's performance. However, we believe that it may have suffered from the same problem experienced by ALPS: its input data were inadequate to describe the meteorological situation in enough detail to permit skilled forecasts to be generated.

GOPAD. GOPAD was hurt by the limitations in the training dataset. This was particularly acute for significant weather. The effects of the presumed underreporting of significant weather in the training data are twofold. First, the climatology calculated from the training data is incorrect, thus leading to the bias shown by GOPAD. Second, and more serious, the set of potentially analogous cases that GOPAD used to develop its set of forecast models was incomplete. This had the effect of limiting the resolving power of the model. Both these effects are evident in the Shootout-89 results: for both correlation (resolving power) and bias, GOPAD performs better for severe weather than for significant.

ALPS. ALPS's poor performance was surprising to the psychologists who developed it because a number of studies have shown that simple linear models of the type used by ALPS perform as well or better than human experts. We hypothesize that the failure of ALPS lies in the selection of cues, not in the way they were aggregated to produce a forecast. Our hypothesis is supported by the better performance of GOPAD and OCI, which relied heavily on linear models but used different cues. Our hypothesis could be tested by using the input variables to other models as cues in a linear prediction model.

A post-facto attempt to fit a linear model using only ALPS's cues to the data was unsuccessful. Although the variables used by ALPS are generally recognized to be important factors in severe and significant weather, we conclude that the individual pieces of data used to determine the cue values were insufficiently representative of the meteorological situation as a whole to allow adequate forecasting.

9. Lessons learned and plans for the future

We learned several lessons from Shootout-89 that will be applied to similar experiments in the future.

- A balance must be struck in choosing the forecast task, so that skill scores are high enough to allow us to discriminate among systems, while maintaining as high a potential utility as possible. Thus, in the future, we will
  1) forecast slightly later in the day, when the boundary layer is more representative of the overlying air mass, yet early enough that the forecast task remains challenging and useful.
  2) forecast for larger areas, as well as for regions of the current size, in order to assess the spatial sensitivity of the systems.
- Generate compatible forecasts made by humans not aided by Shootout systems to put the system forecasts in perspective.
- For systems that require subjective input, test sensitivity to that input by having several meteorologists run the systems independently for a large number of cases.
- Take considerably more care in calibrating the systems and use a better historical training set.

We are currently planning a Shootout-91 experiment. From mid-March through May 1991, systems operating from Boulder will forecast for the region around Norman, Oklahoma. From mid-May through mid-August, the systems will forecast for northeastern Colorado. In both locations, human forecasters will also generate compatible forecasts. In addition, a mesoscale numerical model will also participate for Colorado. We expect that the real fruits of Shootout-89 will be seen in the improved systems that will participate in Shootout-91.

Acknowledgments. The authors would like to thank FSL's Analysis and Prediction Program and PROFS for their support of Shootout-89. In particular, we greatly appreciate the efforts of Darien Davis and Mark Govett, who provided computer support, and Robert Lipschutz, who provided NEXRAD radar products. We would also like to thank the NWS's Denver office for use of their observer networks for verification data. We thank John Brown for painstakingly analyzing each of the many critical wind events, and Jeff Torgerson for reviewing all the verification information. Professor Lew Harvey of the University of Colorado graciously performed the signal detection theory analysis. Funding provided by the director of NOAA's Environmental Research Laboratories, Dr. J. O. Fletcher, by Environment Canada's Atmospheric Environment Service, and by Consultant's Choice, Inc., is greatly appreciated. We thank Hewlett Packard, Inc., for donating the computer hardware used by KASSPr. This material is based in part on work supported by the National Science Foundation under Grant No. SES-8822319.

References

Bullas, J., J.C. McLeod and B. de Lorenzis, 1990: Knowledge Augmented Severe Storms Predictor (KASSPr)—An operational test. *Preprints: 16th Conference on Severe Local Storms,* Boston, Amer. Meteor. Soc., 106–111.

Dawes, R.M., D. Faust and P.E. Meehl, 1989: Clinical versus actuarial judgment. *Science,* **243,** 1668–1674.

de Lorenzis, B., 1988: Interactive graphics editor. *Preprints: 4th International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology,* Boston, Amer. Meteor. Soc., 143–145.

Donaldson, R.J., Jr., R.M. Dyer and M.J. Kraus, 1975: An objective evaluator of techniques for predicting severe weather elements. *Preprints: 9th Conference on Severe Local Storms,* Boston, Amer. Meteor. Soc., 321–326.

Dorfman, D.D., and E. Alf, Jr., 1969: Maximum-likelihood estimation of signal detection theory and determination of confidence intervals—Rating method data. *J. of Math. Psych.*, **6,** 487–496.

Heideman, K.F., 1989: Evaluation of PROFS 1987 convective weather forecasts. *Preprints: 11th Conference on Probability and Statistics in Atmospheric Sciences,* Boston, Amer. Meteor. Soc., 156–160.

Klitch, M.A., J.F. Weaver, F.P. Kelly and T.H. VonderHaar, 1985: Convective cloud climatologies constructed from satellite imagery. *Mon. Wea. Rev.,* **113,** 326–337.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

McGinley, J.A., S.C. Albers and P.A. Stamus, 1991: Validation of a composite convective index as defined by a real-time local analysis system. *Wea. Forecasting,* **6,** 337–356.

Miller, R.G., 1962: Statistical prediction by discriminant analysis. *Meteor. Monogr.,* **No. 4,** 54 pp.

Moninger, W.R., D.F. Cote, J. Davis, R. Dyer, R. Kittredge, R. McArthur, A.H. Murphy and I.R. Racer, 1987: Summary of the First Conference on Artificial Intelligence Research in Environmental Science (AIRIES). *Bull. Amer. Meteor. Soc.,* **68,** 793–800.

Swets, J.A., 1988: Measuring the accuracy of diagnostic systems. *Science,* **240,** 1285–1293.

Weaver, J.F., and R.S. Phillips, 1987: Mesoscale thunderstorm forecasting using RAOB data, surface mesonet observations, and an expert system shell. *Preprints: Symposium on Mesoscale Analysis and Forecasting Incorporating "Nowcasting."* Boston, Amer. Meteor. Soc., 327–331.

——, F.P. Kelly, M. Klitch and T. VonderHaar, 1987: Cloud climatologies constructed from satellite imagery. *Preprints: Third International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology.* Boston, Amer. Meteor. Soc., 169–172.

Zubrick, S.M., and C.E. Riese, 1985: An expert system to aid in severe thunderstorm forecasting. *Preprints: 14th Conference on Severe Local Storms,* Boston, Amer. Meteor. Soc., 117–122.

AMERICAN
METEOROLOGICAL
SOCIETY

This landmark volume, the work of more than 200 contributors, brings alive the remarkable achievements of the past four and one-half decades of radar meteorology. Organized into three segments, the book covers the history, current status, and expected developments in the technological, scientific, and operational subject areas. This is much more than a narrow treatise on radar meteorology. Although the technological chapters deal explicitly with various aspects of the instruments, remaining chapters discuss a broad spectrum of scientific and operational problems such as cloud microphysics and dynamics to synoptic meteorology, nowcasting, severe-storm detection and warning, and hazards to aviation such as low level wind shear. *Radar in Meteorology,* a product of the entire meteorological community, represents an unprecedented compendium of knowledge in the field.

**Edited by
David Atlas**

**RADAR
IN METEOROLOGY**

1990, 806 pages, 420 illustrations/B&W and color, $81 AMS Members/$101 List (clothbound). Index available in IBM 5.25" and 3.5" diskettes, or Macintosh 3.5" diskettes for $30 each. Prices include shipping and handling.Send prepaid orders to: Publications Department, American Meteorological Society, 45 Beacon Street, Boston, MA 02108-3693.