Reprinted from WEATHER AND FORECASTING, Vol. 4, No. 1, March 1989
American Meteorological Society

Analysis of Expert Judgment in a Hail Forecasting Experiment

THOMAS R. STEWART, WILLIAM R. MONINGER, JANET GRASSIA,
RAY H. BRADY AND FRANK H. MERREM

Analysis of Expert Judgment in a Hail Forecasting Experiment

THOMAS R. STEWART,[†,**] WILLIAM R. MONINGER,[*] JANET GRASSIA,[†]
RAY H. BRADY[*] AND FRANK H. MERREM[*]

[*] *Environmental Research Laboratories, National Oceanic and Atmospheric Administration, Boulder, Colorado*
[†] *Center for Research on Judgment and Policy, University of Colorado at Boulder, Boulder, Colorado*

(Manuscript received 15 March 1988, in final form 20 December 1988)

ABSTRACT

This study compared meteorologists, an expert system, and simple weighted-sum models in a limited-information hail forecasting experiment. It was found that forecasts made by meteorologists were closely approximated by an additive model, and that the model captured most of their forecasting skill. Furthermore, the additive model approximated the meteorologists' forecasts better than the expert system did. Results of this study are consistent with the results of extensive psychological research on judgment and decision making processes. Potential implications are discussed.

1. Introduction

The future in weather forecasting is a partnership between person and machine (Snellman 1977; Schlatter 1985; Tennekes 1988), and an understanding of the capabilities and limitations of both is critical to making that partnership effective. Although computer models and algorithms help aggregate weather information for operational forecasters, the human forecaster remains the primary information processor. While a great deal of effort has been devoted to the development of advanced weather forecasting workstations, there has been little study of how forecasters aggregate the information provided by the workstations. The human information processing system is the least understood, yet probably the most important, component of forecasting accuracy.

Human information processing has been a major topic of study by psychologists and others interested in judgment and decision making, and that research has produced a substantial body of knowledge, theories, and techniques that are relevant to the design and implementation of person–machine weather forecasting systems. Three major conclusions drawn from judgment and decision research may have particular relevance for weather forecasting: 1) the results of systematic studies of human information processing yield insights into this process that often contradict people's introspective observations; 2) human information processing is limited and subject to systematic errors and biases; and 3) cognitive assistance can overcome some of the limitations of the judgment process and improve the quality of judgment. For reviews of the research, see Einhorn and Hogarth (1981), Hammond et al. (1980), Hogarth (1980), Sjoberg (1982), Slovic and Lichtenstein (1973), and Slovic et al. (1977).

In this paper we describe an experiment which illustrates how research techniques that have been used by psychologists for over 30 yr can be used to study information processing by weather forecasters. The next section explains how this experiment fits into an overall strategy for investigating the cognitive processes of weather forecasters. Then we describe the experiment, present the results, and discuss the implications.

2. Overview of research strategy

The cognitive processes used in weather forecasting can be divided into three categories: information acquisition, information integration, and output (see Hogarth 1980). Information acquisition is the process of obtaining the information about past and current weather. Each feature of past and current weather (e.g., radar signatures such as reflectivity, rotation, tilt) is a "cue" for the forecast of future weather. Information integration is the activity of assimilating and organizing the cues into a judgment, or set of judgments, about future weather. Output is the process of formulating the forecast into its final form to be issued to the public.

In cognitive psychology, as in most other areas of research, it is necessary to simplify a phenomenon in order to study it. In the present study, we chose to simplify by excluding the perceptual processes involved

[**] Present affiliation: Center for Policy Research, The University at Albany, State University of New York, Albany, New York.

Corresponding author address: Dr. Thomas R. Stewart, Center for Policy Research, Milne 300, The State University of New York—Albany, Albany, New York 12222.

in information acquisition and limiting the forecasters' cognitive activity to information integration and output. As a result, this study concerns only the integration of information to form a forecast, not the perceptual processes involved in acquiring information. Our method (described below) assured that all forecasters in the study used exactly the same information. Consequently, some aspects of forecast skill were necessarily excluded from the study, and a somewhat unrealistic forecasting situation was created because the meteorologists were not able to acquire information as they would in an operational setting.

In the study of complex cognitive processes, there is an inherent trade-off between realism and control that gives rise to a difficult dilemma. We can study cognitive processes in highly realistic situations (e.g., operational forecasters making actual forecasts) where we have very little control, and are therefore not able to draw strong conclusions about the results, or we can conduct controlled studies by introducing constraints (as we did in the present study) so that we can be clear about the results of the experiment, at the expense of introducing doubt about the generality of the results.

The resolution of this dilemma is to include studies representing various points on the realism/control continuum in a research program. When the results of controlled studies are consistent with what is observed in natural settings, we can be confident in our findings. The study to be reported here falls near the low realism/high control end of the continuum. As a result, we can expect to draw relatively clear conclusions about how forecasters integrate information in the experiment ("internal validity") but we must be cautious in generalizing to the cognitive activity of forecasters in operational settings ("external validity"). Despite their limitations, such simplified studies of judgment and decision making have provided important insights into the nature of human cognition (Brown 1972; Kirwan et al. 1983; Dawes 1986). When they are combined with results of more realistic studies (which we have currently planned) the generality of the results can be systematically investigated. Furthermore, when the results of a limited study are consistent with a larger body of theory and research, confidence in generalizations increases. Thus, this study should be viewed as an initial step in the systematic study of human information processing in weather forecasting.

3. Method

Information derived from Doppler radar volume scans of 75 storms was presented to seven meteorologists who then made probability forecasts of hail and severe hail. Two different models, representing alternative ways of describing the meteorologists' subjective judgment processes, were compared with the forecasts. The radar volume scan data, the procedure for obtaining forecasts, and the models are described below.

a. Data

The raw data for the study consisted of 644 Doppler radar volume scans of 156 storms. The data were collected in the summer of 1985 during a forecasting exercise (Haugen 1986) conducted by NOAA's Program for Regional Observing and Forecasting Services (PROFS). The radar was operated by the National Center for Atmospheric Research (NCAR). This radar (CP-2) produced volume scans of reflectivity, Doppler velocity, and differential reflectivity every 5 min, but scans included in the dataset were separated by 10-min intervals. The cues used were determined as part of an earlier project to develop an expert system for hailstorm diagnosis (Merrem and Brady 1988). For that study, a meteorologist (RHB) played back the radar data, and then visually estimated seven cues. The cues were maximum reflectivity at 1) low, and 2) middle levels of the storm, 3) maximum echo gradient within the storm, 4) rotation or convergence within the storm, and 5) tilt of the storm between low and middle levels. The optional cues, which were available for only some of the radar data, were 6) hail signature based on differential reflectivity (ZDR) and 7) upper-level divergence. The severity of each storm was determined from the logs of PROFS chase teams who observed the storms in situ, or from public reports telephoned to the local National Weather Service office. It was necessary to modify the original dataset because data were missing in many volume scans, and only volume scans with complete data could be used in this study. Therefore, upper-level divergence information was not used because it was missing in 67% of the volume scans. In addition, 191 volume scans were dropped because the ZDR signature was not available. The dataset used in this study consisted of six cue variables for the remaining 453 volume scans. Examination of these cases showed they were similar to the original set. The cues and the scoring criteria are listed below.

1) *Reflectivity of core at low level.* From the low-level (0.7 deg) reflectivity PPI scan, estimate the average reflectivity of the storm's core, assuming it consists of at least seven–ten pixels. (Note: In the summer of 1985, a pixel of data displayed on the monitors of the PROFS workstation corresponded to a 500 m × 500 m square.)

2) *Reflectivity of core at middle level.* From the middle-level (6.4 km AGL) reflectivity CAPPI (constant altitude) scan, estimate the average reflectivity of the storm's core, assuming it also consists of at least seven–ten pixels.

3) *Strong echo gradient.* Is there an area of echo (i) at low or middle-levels, (ii) a few kilometers or more in length, and (iii) situated on the SE, S, SW, or advancing flank of the storm where the reflectivity gradient exceeds 8 dBZ km^{-1}?

4) *Tilt.* Comparing the middle-level CAPPI and low-level PPI scans, (i) Is the middle-level high reflectivity core situated over the strong low-level reflectivity

gradient? or (ii) does a horizontal distance of approximately 4 km or more separate the centers of the two cores?

5) *Rotation.* In terms of velocity difference, what is the magnitude of the strongest (cyclonic or anticyclonic) shear or convergence signal observed within the echo at either low or middle levels?

6) *Favorable ZDR signature.* Do the low-level (0.2 deg) differential reflectivity data show a coherent (several pixels) hail signal with this cell?

Verification. In the set of 453 volume scans, either significant (diameter ⩾ 0.25 in. or small hail ⩾ 1 in. deep) or severe (diameter ⩾ ¾ in.) hail was verified within 30 min after 16.1% of the observations, and severe hail was verified after 6.6% of the observations. The problems associated with the verification of severe weather events have been discussed by Hales (1987). Severe storms which track across densely populated urban areas are more likely to be verified as such than are severe storms which remain over sparsely populated rural areas. Although potentially severe storms occurring over rural areas generally had a PROFS chase team assigned to them, it is likely that some of the significant or severe hail events accompanying these storms were not observed by chase teams. In addition, all hail reports were strictly interpreted; i.e., a storm reported as producing hail at 1539 LST was not assumed to be a hail producer at 1540 LST unless it was reported as hailing at the later time. Even though the majority of potential hail-producing storms were observed by chase teams, a few storms were undoubtedly missed. Although we consider our verification dataset to be one of the most complete ever assembled during a real-time forecast experiment, these inherent problems remain.

b. The forecasts

Seven meteorologists made 30-min probabilistic hail forecasts for a sample of 75 volume scans drawn from the original 453. The participants were all research meteorologists who had participated in one or more real-time forecasting experiments using the PROFS workstation. A stratified random sampling procedure was used to select the 75 volume scans to ensure that the base rate (proportion of volume scans for which hail was verified) in the sample matched that in the population of 453 volume scans. Because an error was discovered in the verification data after the study was run, the base rate in the sample turned out to be 14.7% for significant or severe hail and 5.3% for severe hail only.

On the basis of the six cue variables for each volume scan, the meteorologists estimated probabilities both for any hail (significant or severe) and for severe hail only. Figure 1 illustrates how the volume scans were presented to the meteorologists. For reasons described in section 2, the levels of the cues for each volume scan were specified; i.e., meteorologists did not perceive them directly from the radar display as they would in operational forecasting.

The meteorologists expressed concern about the limited information they were given. They said that to forecast hail they would need additional information, for example, about the evolution of the storm, the storm's relation to the surrounding environment, and its location relative to the radar. We explained that the information provided was determined by the availability of data and that we recognized that forecasting skill exhibited in this study could be substantially different from the skill of forecasters in the field.

The 75 volume scans were presented in random order. After judging the first 50 volume scans, participants took a brief break and then judged the remaining 25 volume scans plus an additional 25 volume scans consisting of the even-numbered volume scans from the first set of 50, presented in random order. Repetition of 25 volume scans makes it possible to assess the consistency of the forecasts. No meteorologist reported noticing the repeated volume scans. All meteorologists evaluated 100 volume scans and filled out a questionnaire about their forecasting strategy in less than 2 h.

c. The models

Cognitive processes can be studied in the same way that other natural processes are studied, i.e., by developing alternative models and evaluating those models. Two information processing models were used in this study, and they were evaluated with regard to two criteria: 1) How well does the model reproduce the judgments of the meteorologists? and 2) How well does the model capture forecasting skill? i.e., How accurately does it forecast hail probability? Each model is described below.

1) MULTIPLE REGRESSION

A technique called "judgment analysis," which uses multiple regression analysis to model the judgments of experts, has been used extensively in psychology (Hammond et al. 1975; Stewart 1988). The effectiveness of this technique is based on a pervasive finding in research on judgment and decision making: in many domains of expertise, simple algebraic models can be used to reproduce the judgments of experts (Slovic and Lichtenstein 1973). Often a simple linear model works as well as or better than more complex models (Dawes and Corrigan 1974).

Using judgment analysis, models of the following form were statistically fit to the forecasts made by each meteorologist:

$$Y_{ij} = c_j + b_{j1}X_{i1} + b_{j1}^*(X_{i1})^2 + b_{j2}X_{i2} + b_{j2}^*(X_{i2})^2 + b_{j3}X_{i3} + b_{j4}X_{i4} + b_{j5}X_{i5} + b_{j5}^*(X_{i5})^2 + b_{j6}X_{i6} + e_{ij},$$

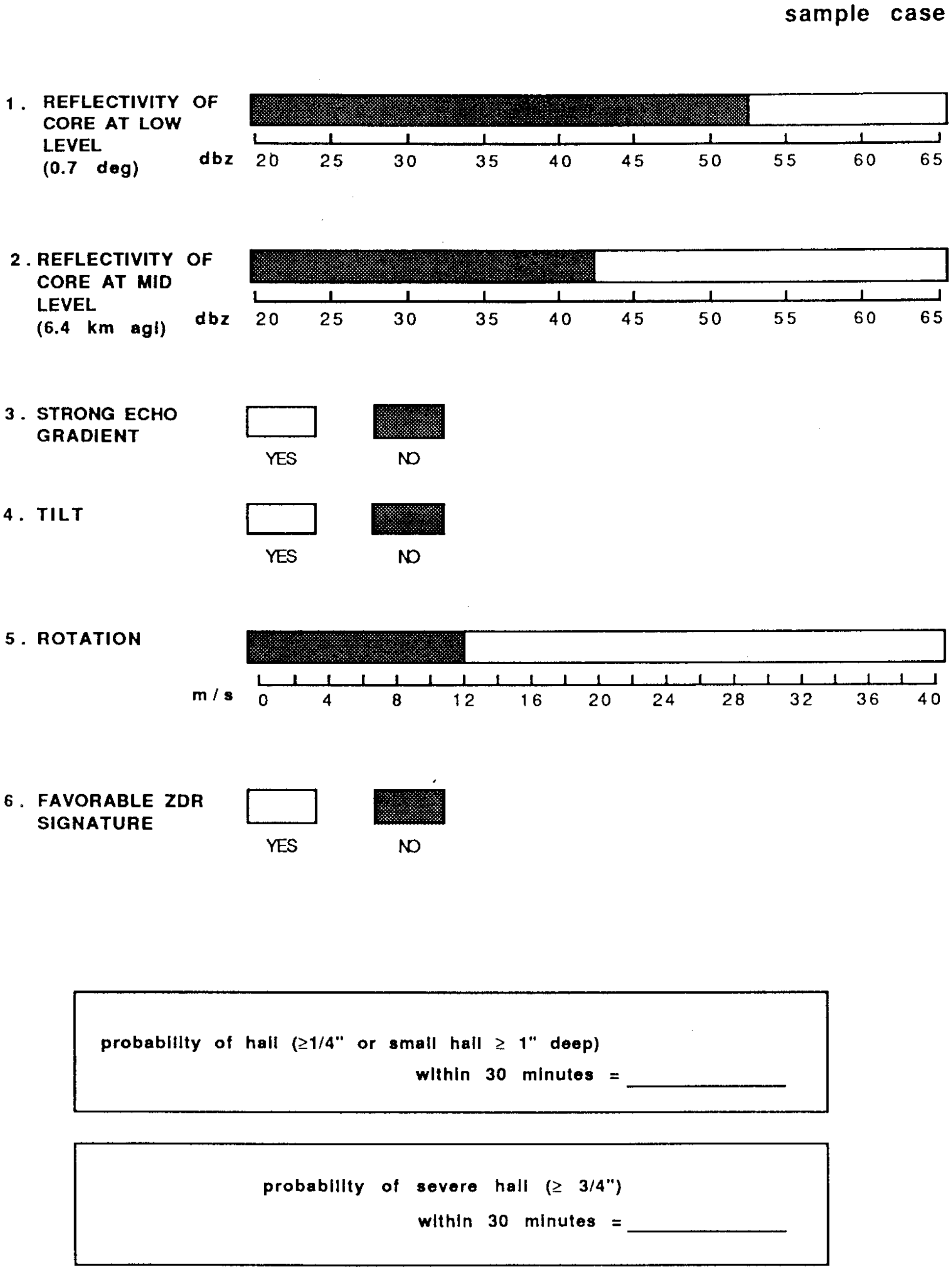
sample case
1. REFLECTIVITY OF CORE AT LOW LEVEL (0.7 deg)
dbz
20 25 30 35 40 45 50 55 60 65
2. REFLECTIVITY OF CORE AT MID LEVEL (6.4 km agl)
dbz
20 25 30 35 40 45 50 55 60 65
3. STRONG ECHO GRADIENT
YES
NO
4. TILT
YES
NO
5. ROTATION
m/s
0 4 8 12 16 20 24 28 32 36 40
6. FAVORABLE ZDR SIGNATURE
YES
NO
probability of hail (≥1/4" or small hail ≥ 1" deep) within 30 minutes = ______
probability of severe hail (≥ 3/4") within 30 minutes = ______

FIG. 1. Sample of a representation of a volume scan.

where

Y_{ij} the forecast made by meteorologist j based on volume scan i
c_j a constant for meteorologist j
b_{jk} the weight for cue k
b_{jk}^* the weight for the square of cue k
X_{i1} the low-level reflectivity for volume scan i
X_{i2} the middle-level reflectivity for volume scan i
X_{i3} the strong echo gradient for volume scan i (0 = no, 1 = yes)
X_{i4} the tilt for volume scan i (0 = no, 1 = yes)
X_{i5} the rotation for volume scan i
X_{i6} the ZDR for volume scan i (0 = no, 1 = yes) and
e_{ij} the residual for meteorologist j on volume scan i

The parameters (c_j, b_{jk}'s and b_{jk}^*'s) of the model were determined so that the sum of the squared differences between the predictions of the model and the actual forecasts were a minimum; that is, for meteorologist j, the sum of the $(e_{ij})^2$ over all of the cases is minimized.

The squares of low- and midlevel reflectivity and rotation were included in the model because plots of the meteorologists' judgments vs these cues suggested that most meteorologists used them in a nonlinear fashion, particularly when they judged the probability of severe hail. The plots indicated that, in many cases, the slope of the curve relating probability forecasts to cue values increases as the cue increases, as if the meteorologists were using the cues exponentially. This occurred much more frequently for the low- and middle-level reflectivity cues than for rotation. This may reflect meteorologists' awareness that dBZ, the measure of reflectivity, is a logarithmic scale. The quadratic approximation to the exponential was used because, in an additive model, the use of exponential transformations of the cues results in a statistically intractable model.

The correspondence between the statistical model and the actual forecasts is given by the multiple correlation (R), which can range from 0 to 1, with 1 indicating perfect fit. The squared multiple correlation (R^2) indicates the proportion of variance of the forecasts that is accounted for by the model.

2) Expert system

The goal of research on expert systems has been the development of computer programs that can emulate the behavior of experts. Expert systems contain a knowledge base that can be thought of as a model of how the expert aggregates information. Thus, an expert system is a model of human information processing. For reviews of expert systems research, see Waterman (1986) or Winston (1984). The relation between research on expert systems and judgment and decision research has been discussed by Hammond (1987a), Stewart and McMillan (1987), and Carroll (1987).

An expert system called HAIL, developed by Merrem and Brady (1988), was used in this study. HAIL consists of 250 rules based on the seven cue variables described in section 3a. Input to the system is provided by an experienced meteorologist. Output consists of statements ordered from 1 to 5 (see Table 1). In addition to diagnosing the presence of hail, the system provides information about the possibility of tornadoes and strong winds. As is typical of expert systems, the 250 rules were derived by discussion with only one person. The rules were designed to represent as closely as possible the thinking process used by the chosen expert meteorologist as he diagnoses storm severity. Since development of an expert system is extremely time consuming, it was not possible to develop one for the other meteorologists in the experiment.

Since the meteorologists made 30-min probability forecasts whereas HAIL was designed to provide categorical diagnoses of hailstorms, it was necessary to transform the output of HAIL so that it could be compared with the probability forecasts. This transformation was accomplished by computing the relative frequency, in the original 453 volume scans, of hail or severe hail within 30 min, given each categorical output (Table 1). These relative frequencies, which are estimates of the conditional probability of hail given the diagnosis, were substituted for the categorical diagnoses. In other words, the output of HAIL was calibrated with respect to the 453 volume scans in the original dataset, and thus was converted from categorical diagnoses into probability forecasts. This procedure makes it possible to validate HAIL's forecasts as probability forecasts (Murphy 1986).

TABLE 1. Calibration of the HAIL expert system.

Diagnosis category*	Number of times given	Number of occurrences of hail within 30 min	Probability of hail, given diagnosis
		Any Hail	
1	251	10	.040
2	60	14	.233
3	75	26	.347
4	34	12	.353
5	33	11	.333
		Severe Hail	
1	251	4	.016
2	60	8	.133
3	75	7	.093
4	34	3	.088
5	33	8	.242

* Description of diagnosis categories: 1) This storm is not significant and not severe. Hail of any size and/or gusty winds are very unlikely. 2) There is a very low probability that this cell may be producing small hail (<¾ in.) and/or moderately strong wind gusts (35–49 kt). 3) This storm is a significant weather producer with small hail (<¾ in.) and/or gusty (35–49 kt) winds. 4) This storm is a significant weather producer wth small hail (<¾ in.) and/or gust (35–49 kt) winds. There is the possibility that it may also be severe with large (≥¾ in.) hail and/or strong (≥50 kt) winds. 5) This storm is severe with large hail (≥¾ in.) and/or strong (≥50 kt) winds.

4. Results

Three types of results are discussed here. First, we describe characteristics of the meteorologists' forecasts. How well do they agree, how consistent are they, and how accurate are they? Then we report on the correspondence between the regression models and the expert-system model and the meteorologists' forecasts. Finally, we compare the accuracy of the meteorologists and the models in order to determine how much of the meteorologists' skill is captured in the models.

a. The meteorologists' forecasts

1) Agreement

Correlations among the seven meteorologists' forecasts (A–G) are presented in Table 2. (Correlations

Table 2. Agreement among meteorologists.

Forecast	A	B	C	D	E	F	G
				Any Hail			
A	(.93)*						
B	.91	(.95)					
C	.86	.83	(.89)				
D	.85	.88	.87	(.95)			
E	.90	.88	.91	.84	(.93)		
F	.84	.88	.75	.79	.77	(.92)	
G	.88	.89	.82	.85	.86	.84	(.95)
	Range .75–.91			Median .86			
				Severe Hail			
A	(.97)						
B	.93	(.96)					
C	.87	.88	(.92)				
D	.84	.90	.95	(.95)			
E	.86	.86	.80	.78	(.69)		
F	.88	.92	.82	.86	.78	(.94)	
G	.87	.90	.84	.85	.92	.85	(.93)
	Range .78–.95			Median .86			

* Numbers in parentheses are estimates of consistencies based on 25 repeated trials.

can range from −1.0 to +1.0.) Agreement among meteorologists was moderate to high for both hail and severe hail forecasts. For forecasts of any type of hail, meteorologist F has the lowest level of agreement with other forecasters, but this is not the case for forecasts of severe hail.

2) Consistency

The numbers in parentheses in the diagonal of Table 2 are estimates of the consistency of each meteorologist's forecasts. A meteorologist who made exactly the same forecasts on repeated presentations of the same information would have a consistency of 1.0. Consistency is estimated by correlating the two sets of judgments of 25 repeated volume scans. The forecasters are not perfectly consistent, but their consistency is generally high except for meteorologist E's forecasts of severe hail. His low consistency is due to a few pairs of repeated volume scans for which he gave two quite different probabilities. In one volume scan, his first forecast was 10% and his second was 50%. If this volume scan were eliminated, his consistency would be 0.82.

3) Performance

Skill scores, squared correlation coefficients, conditional biases, and unconditional biases for each forecaster are presented in Table 3. These indices are described in Murphy (1988). The skill score reported in Table 3 is

$$SS = 1 - [\mathrm{MSE}(f, x)/\mathrm{MSE}(\langle x \rangle, x)],$$

where $\mathrm{MSE}(f, x)$ is the mean square error for the forecast (f) relative to the observed event (x) and $\mathrm{MSE}(\langle x \rangle, x)$ is the mean square error for a constant forecast of $\langle x \rangle$ which is the climatological probability of hail in the sample. This measure reflects the accuracy of the forecasts relative to a reference forecast. The maximum skill score is 1.0, and if the MSE for the forecast is equal to the MSE for the climatological forecast, skill is 0.0.

Squared correlations between forecast probabilities and dichotomous variables representing the occurrence of hail and severe hail (0 = no hail, 1 = hail) are also reported in Table 3. The correlation between a probability forecast and a dichotomous verification variable is a point biserial correlation [see Edwards (1976) for a discussion of the properties of this correlation coefficient] and can range from −1.0 to 1.0. This correlation measures the extent to which forecast probabilities are consistently higher when hail occurs than when it does not. The correlation would be 1.00 if 1) the forecast probability were always p_1 when hail occurred, 2) the forecast probability were always p_2 when hail did not occur, and 3) $p_1 > p_2$, regardless of the values of p_1 and p_2. The correlation will be small when variation in the forecasts, given occurrence or nonoccurrence of hail, is large relative to the total variation in forecasts. It is not sensitive to the actual probabilities or to their range; i.e., a forecaster who always gave probabilities between 0.10 and 0.20 could have the same correlation as another forecaster whose probabilities ranged from 0.50 to 1.00. The correlation measures the ability of the forecast to discriminate consistently between occurrence and nonoccurrence of hail. It does not measure "bias," i.e., the extent to which the magnitudes of the forecast probabilities are appropriate for the weather events being forecast. Two kinds

Table 3. Skill scores, correlation, and bias.

Forecaster	Skill score	Squared correlation	Conditional bias	Unconditional bias
		Forecasts of Any Hail		
A	.046	.233	.079	.108
B	−.340	.181	.114	.408
C	.064	.177	.034	.079
D	−.881	.206	.048	1.039
E	.080	.219	.074	.065
F	−1.018	.125	.331	.811
G	−.704	.154	.264	.594
		Forecasts of Severe Hail		
A	.087	.211	.098	.025
B	−.245	.162	.259	.149
C	−.155	.074	.205	.024
D	−.586	.091	.466	.211
E	−.015	.092	.094	.013
F	−.730	.128	.618	.240
G	−.849	.119	.624	.344

of bias identified by Murphy (1988) are reported in Table 3. "Conditional bias" is related to the slope of the regression line relating observed events to forecasts. Conditional bias is zero only when the slope is 1.0. "Unconditional bias" is related to the difference between the mean forecast and the mean event. It is zero only when these two means are equal. Murphy showed that the skill score is equal to the squared correlation coefficient minus the sum of the two bias terms. He pointed out that since the bias terms cannot be negative, the correlation coefficient might be considered a measure of the "potential skill" that might be attained if all conditional and unconditional biases were eliminated.

Most skill scores in Table 3 are negative and the maximum improvement over climatology is only 8.7%. The correlation coefficients, however, indicate that forecasters were able to distinguish between hail- and nonhail-producing storms to some degree. All correlations were positive and significantly different from 0.0 at the 0.01 level of significance. The low skill scores are due to high levels of conditional and unconditional bias. Thus, Table 3 suggests that meteorologists can *potentially* improve over climatology by more than 20%, but they do not achieve that level of improvement because of biases in the forecast.

b. Models of the meteorologists' forecasts

1) Regression analysis

Table 4 presents squared multiple correlations that have been adjusted to correct for overfitting of the regression model due to the number of predictors relative to the number of volume scans. They indicate that the regression models account for 80%–92% of the variance in the meteorologists' forecasts. In other words, these simple weighted-sum models can reproduce the forecasts with a high degree of accuracy and account for nearly all the consistent variation in forecasts. (See Table 2 for proportion of variance that is consistent for each forecaster.)

This result may seem puzzling because the meteorologists invariably reported that their judgment processes involved nonadditive, synergistic aggregation of information. The ability of the regression model to describe meteorologists' information aggregation processes is consistent, however, with the research on human judgment cited in section 3.

TABLE 4. Adjusted squared multiple correlations for regression models of forecasts.

Forecaster	Any hail	Severe hail
A	.90	.84
B	.92	.91
C	.86	.81
D	.89	.86
E	.89	.80
F	.83	.90
G	.87	.91

TABLE 5. Relative weights of cues.

	Cue[†]					
Forecaster	LDBZ	MDBZ	GRAD	TILT	ROT	ZDR
			Any Hail			
A	.21*	.24*	.13*	.03	.17*	.22*
B	.22*	.27*	.09*	.05	.19*	.17*
C	.17*	.36*	.17*	.08	.12*	.10*
D	.28*	.30*	.10	.14*	.09	.09*
E	.18*	.47*	.07	.04	.10	.14*
F	.30*	.14	.04	.02	.34*	.17*
G	.19*	.42*	.00	.07	.15*	.17*
			Severe Hail			
A	.12	.36*	.21*	.00	.25*	.06
B	.30*	.30*	.08	.01	.22*	.08*
C	.11	.40*	.23*	.09	.07	.10
D	.28*	.28*	.17*	.10*	.13*	.05
E	.14	.68*	.12	.00	.05	.00
F	.26*	.17*	.13*	.00	.33*	.12*
G	.16*	.59*	.00	.04	.15*	.06

* Significant at the .01 level.

[†] LDBZ reflectivity of core at low level; MDBZ reflectivity of core at midlevel; GRAD strong echo gradient (yes, no); TILT tilt (yes, no); ROT rotation or convergence ($\mathrm{m\ s^{-1}}$); ZDR favorable ZDR signature (yes, no).

Regression models can be used to infer how the meteorologists weigh information when they make forecasts. Relative weights of the cues, derived from the regression models, are presented in Table 5 (see the Appendix for derivation of weights). These weights are useful because they can explain, in part, why different meteorologists arrive at different forecasts. In this study, the cues were moderately intercorrelated (Table 6), and, as a result, the weights must be interpreted with caution. The weights that are significantly different from zero (at the 0.01 level of significance) are indicated in the table.

Although the weights differ among meteorologists, they indicate that low- and midlevel reflectivity are generally the most important cues. The notable exception is meteorologist F. For both hail and severe hail, rotation is F's most important cue.

Actual agreement among meteorologists (Table 2) is greater than would be expected based on the differ-

TABLE 6. Cue intercorrelations.

	LDBZ	MDBZ	GRAD	TILT	ROT	ZDR
LDBZ	1.00	.60	.62	.28	.41	.32
MDBZ	.60	1.00	.49	.33	.49	.28
GRAD	.62	.49	1.00	.21	.50	.27
TILT	.28	.33	.21	1.00	.20	.06
ROT	.41	.49	.50	.20	1.00	.19
ZDR	.32	.28	.27	.06	.19	1.00

ences between the weights. This occurs because the cue intercorrelations (Table 6) are all positive. When cues are intercorrelated, different weighting strategies can produce similar forecasts because the cues provide partially redundant information. In this circumstance, agreement among forecasts may be considered "false agreement" (Hammond et al. 1975) because it does not reflect agreement in the underlying forecasting strategy; i.e., there is agreement in fact but not in principle. In the relatively infrequent volume scans when cues diverge, i.e., when some cues indicate hail while other cues indicate no hail, disagreements among meteorologists will emerge. Thus, meteorologists can be expected to disagree most when forecasting is most difficult.

2) THE EXPERT SYSTEM

Correlations between the HAIL expert system and the meteorologists' ranged from 0.70 to 0.85 for forecasts of any hail and from 0.63 to 0.79 for forecasts of severe hail. For all meteorologists, the weighted-sum judgment analysis models reproduced meteorologists' forecasts better than did the HAIL expert system. This includes the forecasts of the meteorologist who developed the rule base for HAIL.

c. Performance of the models

1) REGRESSION MODELS

To what extent do the regression models of the meteorologists capture the accuracy in their forecasts? To answer this question, the regression models described above were applied to the 75 volume scans to produce forecasts. Performance of these models is described in Table 7. Only the correlation coefficients which, as described above, indicate the potential skill of an unbiased forecast, are reported here. In the case of the regression model, unconditional bias of the model is identical to that of the forecaster. Changes in conditional bias reflect changes in the correlation coefficient and in the variance of the forecasts.

TABLE 7. Performance of forecasts and models of forecasts (correlations).

Forecaster	Original forecasts	Regression models
	Any Hail	
A	.48	.41
E	.47	.45
D	.45	.43
B	.43	.42
C	.42	.45
G	.39	.43
F	.35	.37
	Severe Hail	
A	.46	.37
B	.40	.37
F	.36	.35
G	.34	.37
E	.30	.35
D	.30	.34
C	.27	.34

The models capture most of the (potential) skill in the forecasts for six of the seven meteorologists. Only meteorologist A substantially outperforms the model that is based on his judgments.

The rows of Table 7 have been ordered from highest to lowest correlation of the original forecasts to highlight a pattern in the data. For the least accurate meteorologists, the model outperforms the original forecasts; but for the most accurate meteorologists, the model does worse than the original forecasts. Thus, differences in performance among the models are less than the differences among the original forecasts. This suggests that some (small) component of accuracy (or inaccuracy) may not be captured by the regression models. Whether that component is simply chance (lucky or unlucky forecasts) or a systematic, synergistic process remains to be determined in further research.

The small differences among the correlation coefficients for different regression models in Table 7 also reflect a "flat maximum" effect (Lovie and Lovie 1986; von Winterfeldt and Edwards 1982) due to intercorrelations among the cues. When cues are intercorrelated, it may not matter much how the information provided is integrated into a forecast as long as it is done in a reasonable and consistent fashion. In the hail data used in this study, the cues were intercorrelated (Table 6), the relations between the cues and the probability of hail were all monotonic, and, given the data provided, there was a high degree of uncertainty about whether a storm would produce hail. These are all contributing factors to the flat maximum effect.

For any task with these properties, a weighted-sum model will perform about as well as any other model, and the magnitudes of the weights do not matter much as long as they have the correct sign (Dawes and Corrigan 1974). Researchers have found that the weighted-sum model generally outperforms humans for these kinds of tasks because the model is perfectly consistent whereas the human is not (Goldberg 1969, 1970; Camerer 1981). The model proves superior even though it does not include complex interactions among the cues, or "synergisms," which are important to human experts.

2) EXPERT SYSTEM

For forecasts of any hail, the correlation for HAIL is 0.38, slightly above the lowest correlation for a meteorologist. For severe hail forecasts, the correlation for HAIL is 0.41, which is near the level of the best meteorologist and slightly better than his regression model.

5. Discussion

This study illustrated how the subjective component of forecasting can be systematically studied. The design of the experiment made it possible to investigate the following characteristics of the forecasts:

- *Agreement.* Agreement among forecasts was moderately high in this study. Lack of agreement (see Lusk et al. 1988, for example) may indicate that some forecasters are inconsistent or that they are using different forecasting strategies.
- *Consistency.* If the forecasting process is consistent, then identical conditions produce identical forecasts. If the forecasting process is not consistent, then there is a degree of arbitrariness about the forecasts that will reduce their accuracy. In this simple experiment, the forecasts were highly consistent. In general, as the amount of information and the complexity of a task increases, consistency decreases. This fact suggests that forecasts in the field may be less consistent than those in this experiment.
- *Descriptive model.* Statistical regression models provided good descriptions of the forecasts. Furthermore, the regression models were generally as accurate as the original forecasts. In comparison with a complex expert-system model, the regression models provided better approximations to the meteorologists' forecasts and were just as accurate.
- *Parameters of judgment models.* It is useful to describe judgment processes in terms of weight, function form, and organizing principle (Hammond et al. 1975). Weights reflect the relative importance of different items of information. The weights estimated in this study (Table 5) indicated that different meteorologists attached different importance to the cues. Function forms describe the relation between each cue and the forecast. In this study, the reflectivity cues and rotation were related to the forecasts by an exponential function form. The organizing principle governs the way that the various cues are organized into an overall forecast. The organizing principle implicit in the regression models is additive. The expert system employs a nonadditive, synergistic organizing principle. In this study, the additive organizing principle provided the best approximation to the meteorologists' forecasts.

Further research is needed to determine the generality of the results found in this study. In particular, studies involving more realistic forecasting situations are necessary. It must be stressed, however, that our results are consistent with a large body of research and theory in judgment and decision making. It is likely, therefore, that they can be applied to some situations that arise in operational forecasting.

6. Conclusion

The importance of studying the subjective judgment processes involved in weather forecasting is supported by the work of Allen (1982), Allen et al. (1986), and Allan Murphy and his colleagues (e.g., Murphy and Winkler 1971; Murphy and Brown 1984). Our study has shown that research methods used by psychologists to study human judgment processes can be applied to weather forecasting. The experiment suggests that the intuitive processes that weather forecasters use to aggregate information into a forecast can be analyzed and described in quantitative terms.

A number of interesting and important forecasting questions can be addressed using systematic methods borrowed from judgment and decision research. For example, how do novice and experienced forecasters differ with regard to consistency, relative weights, function forms, and organizing principle? What is the effect of advanced workstations on the forecaster's judgment processes? Does additional information reduce the consistency of forecasts, and, if so, how can consistency be increased? Can feedback about judgment parameters be used to improve forecasting skill (Hammond et al. 1975; Hammond 1987b)? How much of the skill of expert forecasters can be captured by computers?

Continued research on cognitive processes in weather forecasting is likely to prove useful in the design of "person–machine" systems for weather forecasting. Design of such systems must be based on realistic views of both machine and human capabilities. Through research in computer science and artificial intelligence, machine capabilities are being expanded. Through the study of human information processing in weather forecasting, we are gaining an understanding of the human judgment process.

Acknowledgments. Special thanks are due the expert meteorologists who participated in this study. C. Bullock, E. Szoke, and L. Murdock provided helpful comments on an earlier version of this paper. We are indebted to K. Hammond, J. McGinley, and S. MacDonald for their support and encouragement. We also wish to thank D. Walker and M. McCoy for providing advice on forecast evaluation and M. Luhring and D. Victor for preparing the manuscript.

This research was supported by the Office of Basic Research, Army Research Institute, Contract MDA903-86-C-0142 and by the following branches of the National Oceanic and Atmospheric Administration: the Artificial Intelligence Program; the Program for Regional Observing and Forecasting Services, Environmental Research Laboratories; and the Office of Meteorology, National Weather Service. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

APPENDIX

Calculation of Relative Weights

Regression weights do not indicate the relative importance of the cues because (a) the cues are expressed in different units and (b) there are two weights for each of the continuous cues because the squared terms were included in the regression analysis. The following procedure was used to calculate the relative weights listed in Table 5.

1) For each forecast, regression weights for the model described in section 3c were computed.

2) For forecast j, the continuous cues (low-level reflectivity, midlevel reflectivity, and rotation) were transformed as follows:

$$f_k(X_{ik}) = b_{jk}X_{ik} + b_{jk}^{*}X_{ik}^{2},$$
$$\text{for} \quad i = 1\text{–}75 \quad \text{and} \quad k = 1, 2, 5.$$

This transformation combines the two terms for the continuous cues into a single term.

3) A second regression analysis was computed using the three transformed cue variables and the three binary cues to predict the forecast. The regression equation was

$$Y_{ij} = c_j + b_{j1}f_1(X_{i1}) + b_{j2}f_2(X_{i2}) + b_{j3}X_{i3} + b_{j4}X_{i4} + b_{j5}f_5(X_{i5}) + b_{j6}X_{i6} + e_{ij}.$$

This form of the regression equation has only one weight for each cue. It is a simple algebraic transformation of the original regression equation, and the R^2s were identical to those obtained in the original analysis.

4) The regression weights for the standardized form of the regression equation (the beta weights) were summed, and each beta weight was divided by that sum. (The standardized form of the regression equation compensates for differences in units by transforming each variable so that its mean is 0.0 and its variance is 1.0 in the sample.) This calculation gave the relative weights presented in Table 5.

Several methods have been proposed for computing relative weights in judgment analysis. Alternative methods are discussed in Darlington (1968) and Stewart (1988).

REFERENCES

Allen, G., 1982: Probability and judgment in weather forecasting. Preprints, *Ninth Conference on Weather Forecasting and Analysis*, Seattle, Amer. Meteor. Soc., 1-6.

——, V. Ciesielski and W. Bolam, 1986: Evaluation of an expert system to forecast rain in Melbourne. Paper presented at the *First Australian Artificial Intelligence Congress*, Melbourne, 11 pp.

Brown, T. R., 1972: A comparison of judgmental policy equations obtained from human judges under natural and contrived conditions. *Math. Biosci.*, **15**, 205–230.

Camerer, C. F., 1981: General conditions for the success of bootstrapping models. *Organ. Behav. Human Decis.*, **27**, 411–422.

Carroll, B., 1987: Artificial intelligence, Expert systems for clinical diagnosis: Are they worth the effort? *Behav. Sci.*, **32**, 274–292.

Darlington, R. B., 1968: Multiple regression in psychological research and practice. *Psych. Bull.*, **69**, 161–182.

Dawes, R. M., 1986: Representative thinking in clinical judgment. *Clin. Psych. Rev.*, **6**, 425–441.

——, and B. Corrigan, 1974: Linear models in decision making. *Psych. Bull.*, **81**, 95–106.

Edwards, A. L., 1976: *An Introduction to Linear Regression and Correlation.* Freeman, 213 pp.

Einhorn, H., and R. M. Hogarth, 1981: Behavioral decision theory: Processes of judgment and choice. *Ann. Rev. Psych.*, **32**, 53–88.

Goldberg, L. R., 1969: The search for configural relationships in personality assessment: The diagnosis of psychosis vs. neurosis from the MMPI. *Multivariate Behav. Res.*, **4**, 523–536.

——, 1970: Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psych. Bull.*, **73**, 422–432.

Hales, J. E., 1987: An examination of the National Weather Service severe local storm warning program and proposed improvements. NOAA Tech. Memo., NWS NSSFC-15, 32 pp.

Hammond, K. R., 1987a: Toward a unified approach to the study of expert judgment. *Expert Judgment and Expert Systems*, J. L. Mumpower, O. Renn, L. D. Phillips and V. R. R. Uppuluri, Eds., Springer-Verlag, 1–16.

——, 1987b: Annotated bibliography on cognitive feedback. University of Colorado, Center for Research on Judgment and Policy, Tech. Rep. 269, 28 pp.

——, T. R. Stewart, B. Brehmer and D. O. Steinman, 1975: Social judgment theory. *Human Judgment and Decision Processes*, M. F. Kaplan and S. Schwartz, Eds., Academic Press, 271–312.

——, G. H. McClelland and J. Mumpower, 1980: *Human Judgment and Decision Making: Theories, Methods, and Procedures.* Praeger, 258 pp.

Haugen, D. A., 1986: The PROFS RT85 forecast exercise. Reprints, *Eleventh Conference on Weather Forecasting and Analysis*, Kansas City, MO, Amer. Meteor. Soc., 335–339.

Hogarth, R. M., 1980: *Judgement and Choice: The Psychology of Decision.* Wiley, 250 pp.

Kirwan, J. R., D. M. Chaput de Saintonge, C. R. B. Joyce and H. L. F. Currey, 1983: Clinical judgment in rheumatoid arthritis. I: Rheumatologists' opinions and the development of "paper patients". *Ann. Rheum. Dis.*, **42**, 644–647.

Lovie, A. D., and P. Lovie, 1986: The flat maximum effect and linear scoring models for prediction. *J. Forecasting*, **5**, 159–168.

Lusk, C. M., T. R. Stewart and K. R. Hammond, 1988: Judgment and decision making in dynamic tasks: The case of forecasting the microburst. Unpublished manuscript, Center for Research on Judgment and Policy, University of Colorado at Boulder.

Merrem, F. H., and R. H. Brady, 1988: Evaluating an expert system for forecasting. *Proc. Fourth International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Anaheim, Amer. Meteor. Soc., 259–261.

Murphy, A. H., 1986: Comparative evaluation of categorical and probabilistic forecasts: Two alternatives to the traditional approach. *Mon. Wea. Rev.*, **114**, 245–249.

——, 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424.

——, and R. L. Winkler, 1971: Forecasters and probability forecasts: Some current problems. *Bull. Amer. Meteor. Soc.*, **52,** 239–247.

——, and B. G. Brown, 1984: A comparative evaluation of objective and subjective weather forecasts in the United States. *J. Forecasting,* **3,** 369–393.

Schlatter, T. W., 1985: A day in the life of a modern mesoscale forecaster. *ESA Journal,* **9,** 235–256.

Sjoberg, L., 1982: Aided and unaided decision making: Improving intuitive judgment. *J. Forecasting,* **1,** 349–363.

Slovic, P., and S. Lichtenstein, 1973: Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Human Judgment and Social Interaction,* L. Rappoport and D. A. Summers, Eds., Holt, Rinehart & Winston, 16–108.

——, B. Fischhoff and S. Lichtenstein, 1977: Behavioral decision theory. *Ann. Rev. Psych.,* **28,** 1–39.

Snellman, L. W., 1977: Operational forecasting using automated guidance. *Bull. Amer. Meteor. Soc.,* **58,** 1036–1044.

Stewart, T. R., 1988: Judgment analysis: Procedures. *Human Judgment: The Social Judgment Theory View,* B. Brehmer and C. R. B. Joyce, Eds., North-Holland, 519 pp.

——, and C. McMillan, 1987: Descriptive and prescriptive models for judgment and decision making: Implications for knowledge engineering. *Expert Judgment and Expert Systems,* J. L. Mumpower, O. Renn, L. D. Phillips, and V. R. R. Uppuluri, Eds., Springer-Verlag, 305–320.

Tennekes, H., 1988: The outlook: Scattered showers. *Bull. Amer. Meteor. Soc.,* **69,** 368–372.

von Winterfeldt, D., and W. Edwards, 1982: Costs and payoffs in perceptual research. *Psych. Bull.,* **85,** 267–273.

Waterman, D. A., 1986: *A Guide to Expert Systems.* Addison Wesley, 419 pp.

Winston, P. H., 1984: *Artificial Intelligence,* Addison Wesley, 524 pp.