The Weather Information and Skill Experiment (WISE):  
The Effect of Varying Levels of Information on Forecast Skill

KENNETH F. HEIDEMAN  
Phillips Laboratory, Geophysics Directorate, Hanscom Air Force Base, Massachusetts

THOMAS R. STEWART  
Center for Policy Research, University at Albany, State University of New York, Albany, New York

WILLIAM R. MONINGER  
National Oceanic and Atmospheric Administration/Environmental Research Laboratories/Forecast Systems Laboratory, Boulder, Colorado

PATRICIA REAGAN-CIRINCIONE  
University Center for Policy Research, University at Albany, State University of New York, Albany, New York  
(Manuscript received 5 March 1992, in final form 14 September 1992)

ABSTRACT

The relationship between the quality and quantity of information available to meteorologists and the skill of their forecasts was investigated. Twelve meteorologists were asked to make probabilistic forecasts of significant and severe weather events under three information conditions. Forecast accuracy was generally low. As the amount and quality of the information increased substantially, there was a modest increase in the accuracy of the forecasts. However, the results suggest that the forecasters were least consistent when they had the most information to work with, partially reducing the benefits of the increased information.

1. Introduction

We have entered a period in meteorology which is producing significant changes in our way of observing, forecasting, and disseminating information about the atmosphere. Major new technologies developed in the past ten years are expected to result in more accurate and useful weather forecasts. Specifically, Doppler radar, automated surface observing systems, advanced geosynchronous and polar-orbiting satellites, and profilers will allow surveillance of meteorological phenomena, particularly on the mesoscale, to an extent never before possible. A consequence of this will be an unprecedented volume of data for meteorologists to consider before issuing a forecast or warning. As the amount of weather information, scientific knowledge about the atmosphere, and computer power to process information increases, so does the cognitive burden on meteorologists. Perhaps Bosart (1989, p. 271) stated it best: “Over the next couple of years we are going to be flooded with data from new observational systems on smaller spatial and temporal scales. The process of converting all of this data to usable information is going to require considerable human talent.” A deeper understanding of the human element in weather forecasting might well lead to improvements in weather forecasting accuracy through improvements in information displays, forecaster training and provision of information processing aids such as expert systems. Implementation of such methods could help ensure that the full benefits of the new technology ultimately reach the public.

The notion that the assumed benefit of vastly increased amounts of weather information may not yet be fully reflected in operational weather forecasts is the motivation for the Weather Information and Skill Experiment (WISE), an interdisciplinary effort that is part of a research program on human information processing in weather forecasting that was initiated in 1986. That program has involved collaboration between the National Oceanic and Atmospheric Administration (NOAA), the National Center for Atmospheric Research (NCAR), the Center for Research on Judgment and Policy at the University of Colorado, the Center for Policy Research at The State University of New York at Albany, and the Geophysics Directorate of...
Phillips Laboratory at Hanscom Air Force Base in Bedford, Massachusetts. The primary focus of WISE is the relation between the amount and quality of information available to meteorologists and the quality of forecasts they produce given that information.

Human cognitive processes have been extensively studied in fields such as psychology, medicine, law, business, agriculture, and education. For reviews of these studies see Arkes and Hammonds (1986), Hogarth (1987), Slovic et al. (1977), Shanteau and Stewart (1992), Hoffman (1992), and Chi et al. (1988). Because people, including experts, are limited information processors, giving them more information does not necessarily help them. Studies have shown that as certain groups of people were given more information, their confidence in their judgments increased, but their accuracy did not (Arkes et al. 1986; Oskamp 1965). Explanations of the low association between the amount of information available and judgmental accuracy focus on issues surrounding the increasing complexity of the task as more factors are considered. Possible explanations include information “overload,” discrepancies in the way experts integrate information (Brehmer and Brehmer 1988; Einhorn 1974), and inconsistency in the use of information when making judgments (Brehmer and Brehmer 1988; Dawes et al. 1989).

The purpose of WISE is to investigate the extent to which the skill of weather forecasts is affected by the amount and type of available data and by cognitive biases and limitations. More specifically, an attempt is made to determine the applicability to weather forecasting of a conceptual model which is based on judgment and decision-making research (Fig. 1). Simply stated, the model suggests that as the amount of information available to forecasters increases, the maximum forecasting skill attainable, given the event being forecast, increases at a faster rate than the actual skill achieved by forecasters. As a result, the gap between actual and maximum possible skill theoretically widens with increasing information. The following clarifications apply to the model: 1) maximum possible skill is the theoretical limit of skill that can be attained by optimal use of available information. It depends on the quality of information available and on the event being forecast; 2) maximum possible skill cannot decrease with increasing information, but actual skill can decrease; 3) the “cognitive gap” between actual and maximum possible skill can be attributed to “process loss” which can result from i) inappropriate or suboptimal use of information; ii) forecaster bias which may increase as information increases—some research suggests that overconfidence in judgement increases

[Diagram of skill and information relationship]

with increasing information (Stewart et al. 1992a,b); iii) decreasing forecaster consistency with increasing information.

Some of the causes of the cognitive gap, such as forecaster bias and inconsistency, are well known (Murphy and Winkler 1987; Murphy 1988; Stewart 1990). The hope is that research designed to measure the size of the gap and identify the reasons for it can be used as a basis on which to develop specific strategies to reduce their influence in the forecasting process.

A description of the experimental design is presented in the next section. The framework for measuring and decomposing forecast skill is then presented in section 3. This is followed by presentation and discussion of results in section 4. Concluding comments can be found in section 5.

2. Experiment design

Twelve meteorologists took part in this study, including five (in Boulder, Colorado) from the Forecast Systems Laboratory (FSL, formerly the Program for Regional Observing and Forecast Services), NOAA’s Weather Research Program, and NCAR, and seven from the Geophysics Directorate of the Phillips Laboratory at Hanscom Air Force Base in Bedford, Massachusetts. Participants were asked to make 60-minute probability nowcasts of significant and/or severe convection, and tornadoes occurring in northeastern Colorado under three information conditions (see Table 1 for verification criteria pertaining to significant, severe, and tornadic events). In two of the conditions (“map” and “maps with specified cues”), the information was severely limited. In the third, forecasting with full information was simulated using an advanced FSL workstation for Boulder participants; slides of critical products displayed on the FSL workstation were used to satisfy this condition at Hanscom in lieu of a workstation.

The WISE forecast cases were taken from data generated during a real-time convective forecast exercise conducted in northeastern Colorado during the summer

---

1In meteorology, accuracy refers to the match between forecasts and events, while skill refers to the accuracy of forecasts relative to the accuracy of some reference forecast, such as climatology or persistence (Sanders 1963).
of 1987 (RT-87). During each day of the exercise, up to five specially trained chase teams were deployed in the study area to verify the occurrence or nonoccurrence of significant and severe weather. Verification was also obtained from National Weather Service spotters and automated observations from the FSL 22-station mesoscale network (mesonet) which reports temperature, dewpoint, rainfall, wind speed, and wind direction every five minutes. During RT-87, forecasters used data provided by the Denver AWIPS-90 Risk Reduction and Requirements Evaluation, Part I (DARE-I) Workstation at the FSL facility. This state-of-the-art workstation, which incorporated RAMTEK graphics and VAX hardware, was identical to the one used operationally at the Denver Weather Service Forecast Office from 1986 to 1989.

Most of the data available to FSL forecasting teams through DARE-I during RT-87 were archived for later use by scientists within and outside of FSL using the Displaced Real-Time (DRT) capability of DARE-I. This made it possible for individual forecasters to review a given archived case on the workstation at later times. The availability of relatively dense verification data to go along with the archived datasets made the summer of 1987 the preferred source of cases for WISE and limited the focus of the experiment to roughly the RT-87 area of operations (see Fig. 2). Therefore, the level of experience of WISE participants in forecasting convection in northeastern Colorado must be taken into account when results are analyzed. In fact, six of the participants had forecasting experience in the study area (including all five Boulder forecasters), while six did not. Prior to the experiment, those without Colorado forecasting experience were each given several papers describing the convective climatology of the front range. While this was no substitute for experience, it did provide people in this group with a common baseline from which to work.

Only days for which complete archived datasets existed from 1000 to 0100 UTC (0400–1900 local time (LFT)) during the summer of 1987 were considered for use in WISE. This 15-hour window enabled WISE forecasters to assess the preconvective environment and provided data throughout the afternoon and early evening when storm chasers had been on duty and convective activity was most likely.

Due to limited access to the DARE-I workstation and to data storage limitations, the experimental dataset for this study was limited to two individual cases. Two convectively active days for which nearly complete datasets existed were selected: 1) 20 June 1987: 24 significant and 8 severe weather events occurred within the study area. This included 4 tornadoes and multiple reports of hail up to one inch in diameter; 2) 23 June 1987: 41 significant and 26 severe weather events occurred within the study area, many in association with 3 separate rotating-supercell storms. Three tornadoes were reported.

a. Information conditions

The amount of information available to forecasters was varied but other factors affecting the quality of information, such as the kind of information and the way it was displayed, were varied as well. The three information conditions (described below) all had the following properties: (i) the information contained in each level was a subset of the information contained in the next higher level; (ii) in the judgment of expert meteorologists involved in the design of the experiment, each level added new, nonredundant information and should have provided a significantly improved information base for forecasting; (iii) each successive level was designed to provide a better quality of data display. Presentation ranged from simple black and white schematics in the “map” condition to full-color displays of complex datasets with full information. An added dimension for the Boulder forecasters was the ability to easily animate the data displays on the workstation in the full information condition.

b. Map condition

Forecasters had a bare minimum of information to work with in this part of the experiment. Each was given a set of nine maps depicting 50-dBZ storm contours and five-minute tickmarks showing storm tracks corresponding to the centroids of each of the contours during the previous 20 minutes. An example map showing the location of the FSL mesonet stations, the Doppler radar (NCAR's CP-2) used in the experiment, and the twelve forecast areas is shown in Fig. 2. The forecast areas were counties (or in some cases portions of counties) within the CP-2 coverage area and east of the front range in northeast Colorado. Five of the maps were taken from 20 June and four from 23 June (incomplete verification data for one of the hours on 23 June necessitated elimination of the corresponding forecast). Activity on these maps occurred between
2000 and 0000 UTC (1400–1800 LT). The storm contours and tracks were drawn from Doppler radar displays by experienced meteorologists. Only storms showing maximum reflectivity greater than 50 dBZ were represented on the maps. Because WISE forecasters would be working with the same cases in the “maps with specified cues” and “full information” conditions, the nine maps were presented in random order, and neither the time nor date corresponding to each map was provided. The random order of nowcast maps was intended to decrease the possibility of forecaster calibration based on feedback.

Based only on the information on the maps, participants made one-hour nowcasts of probability of either significant or severe convection for the areas of each county that were not shaded (as in Fig. 2). Because of its small size and location, several forecasters overlooked the small nowcast area at the bottom of the unshaded portion of the map (northern El Paso County) on a number of occasions; it was subsequently dropped from the analysis of results. The remaining 11 county nowcasts for each hour resulted in a total sample of 99 nowcasts for each meteorologist in the map condition. All events in the unshaded portion of
the radar circle (Fig. 2), not just those associated with the storms represented on the maps, were included in the verification data. An average of 1.0–1.5 hours per forecaster was needed to complete this part of the experiment.

c. Maps with specified cues

In this information condition, forecasters were given a new set of the nine maps (in a different random order) featuring the same storms plus a set of “storm profiles” corresponding to each contour on the map. These profiles contained information about selected features of each storm and about the environment around the storm (see Fig. 3 for an example profile). The specific features included in the profiles were selected prior to the experiment by a group of three experienced meteorologists as important cues upon which to base nowcasts of convective activity. Two of these meteorologists were subsequently joined by another, and these three made independent judgments of the feature values that were to appear in each storm’s profile. For several of the features (maximum shear, storm tilt, and stability change) the range of values determined by the three meteorologists was large, indicating that perception of storm features from the full information display can be quite subjective. When the experts disagreed, median cue ratings were used in the storm profiles. Though these cues represented only a very small subset of the data that would normally be available to an operational meteorologist, they collectively provided an intermediate level of information between the map and full information conditions in WISE.

Given this additional information on each storm based on radar imagery and other cues, participants were asked to be more specific with their forecasts; three one-hour nowcasts (probability of significant/severe convection, severe convection, and tornado) were made for each of the 11 valid forecast areas on each map. These three events were not mutually exclusive, that is, summation of probabilities could exceed 100%. In all, 99 sets of nowcasts (a total of 297 nowcasts) were made in the maps with specified cue conditions. This phase of the experiment required an average of about two hours per person to complete. The “map” and “maps with specified cues” conditions are collectively referred to as the limited information conditions.

d. Full information

The five Boulder forecasters took part in the full information condition using all the information available in DRT on the DARE-I workstation. All had considerable experience forecasting convection in Colorado and were proficient in using the workstation. Only five Boulder forecasters were recruited for the experiment because this phase of WISE required several hours of uninterrupted time on DARE-I for each forecaster, and demand for the workstation for other purposes was extremely high. Data storage limitations for DRT cases made it necessary to eliminate some products that had been available on DARE-I during real time. However, none of the eliminated products was judged essential to the forecasting task at hand. Taking numerous daily product updates into account, several thousand products were available to WISE forecasters each day on DARE-I. These included upper-air analyses at five levels, skew T-log p and profiler plots, observations from the FSL mesonet, radar imagery (reflectivity and velocity) at two elevation angles, satellite imagery (visible and infrared), and numerical model output from the National Meteorological Center’s Nested Grid Model and the FSL Mesoscale Analysis and Prediction System (Heideman et al. 1989). The utility of DARE-I went well beyond the various products available; since much of the information was in the form of images (many of which were color coded), perception and pattern recognition were expected to play a greater role in the acquisition of information than in the limited information conditions. Furthermore, use of a workstation allowed the forecasters to search for the information they wanted, overlay different types of information for comparison and integration, zoom in on areas of interest in order to see more detail, and loop information to produce animated displays which reveal patterns over time. In addition, the forecasts were made in time order, so that forecasters could make use of dynamic information about trends and changes in conditions.

In the full information condition, WISE forecasters used DARE-I to make nowcasts for the same times and county areas that were included in the limited information conditions. However, in contrast to the lim-

![Fig. 3. Example storm profile showing characteristics of identified storms and their environments.](image-url)
ited information conditions, the forecasters were given the date of the case for which they were to forecast and each nowcast map indicated the specific one-hour valid period of the corresponding nowcast. There was some concern that the Boulder forecasters would recall specific details of the location and severity of events if given this information, particularly because each case involved weather events in northeast Colorado only two summers earlier, and each of the full information forecasters had actually been working in Boulder at that time. Experiment participants in Boulder were therefore asked to complete a “WISE Weather Recall Form” before, during, and after they worked through each case. In completing the form, forecasters were asked to indicate whether and to what extent they had mental recall of the important weather events (or lack thereof) of the case they were working through. None of them remembered any of the cases.

The exercise itself was divided into two parts. In the first part, forecasters reviewed data available in DRT from 1000 to 1829 UTC for each case day in order to assess the preconvective environment and the potential for significant and severe weather later in the day. Forecasters were able to move forward or backward in time, but were allowed a maximum of two hours real time to review the data before submitting a written convective outlook. In the second part of the exercise, forecasters made a sequence of one-hour nowcasts at five different times (2000, 2100, 2200, 2300, and 0000 UTC for 20 June and 2000, 2100, 2200, and 2300 UTC for 23 June), for the probability of 1) significant/severe convection, 2) severe convection, and 3) tornadoes occurring in each of the county areas (see Fig. 2). A total of 99 sets of nowcasts was made for the two case days. No time limit was set for submission of each nowcast, but forecasters were only permitted to evaluate data on the workstation for the hour leading up to the issuance of each nowcast. For example, forecasters could look at data exclusively between 2000 and 2059 UTC for a nowcast valid from 2100 UTC to 2200 UTC. This eliminated the possibility of forecasters seeing what actually transpired during the valid period of their nowcasts prior to issuing the nowcasts. Once a nowcast was submitted, it could not be returned to the forecaster. All five Boulder forecasters completed three information conditions in the following order: 1) full information, 2) map, 3) maps with specified cues. The full information condition was completed first because the availability of the workstation was greatest early in the experiment.

e. The “slide” condition

The FSL workstation was not available to participants outside of Boulder. However, in an attempt to run an analog to the full information condition for such participants, slides were taken of many products displayed on the DARE-I workstation on 20 June and 23 June. The seven participants at Hanscom AFB were given these slides and asked to make nowcasts for the same times and county areas that were included in the limited information conditions (a total of 99 sets of forecasts). Only one of this group had experience (albeit limited) forecasting convection in Colorado. As with the full information condition in Boulder, the slide condition was divided into two parts for each case. In the first part, forecasters were given approximately 50 slides of a wide range of products depicting the preconvective environment, including skewT–logp and upper-air plots, satellite imagery, and numerical model products. In the second part, approximately 30 slides were available each hour and were generally limited to satellite and radar imagery, and mesonet data. Forecasters were able to manipulate the slides as they wished (forward or backward in time) up to the start of the nowcast valid time. Of course, the slide condition could not replicate many workstation features, such as animation, zooming, and resolution, nor could it provide many of the products available on DARE-I. However, the slides collectively provided a cross section of fundamental products crucial to severe weather forecasts, including radar reflectivity and velocity, satellite imagery, mesonet data, etc., and they provided far more information than was available in the limited information conditions. All seven Hanscom forecasters completed three information conditions in the following order: 1) map, 2) maps with specified cues, 3) slide.

3. Measures of skill

Numerous measures of skill and accuracy of forecasts are available (see, for example, Murphy and Winkler 1987, and Stanski et al. 1989). As Murphy and Winkler noted, all the information about forecast accuracy is contained in the joint distribution of forecasts and events, and different measures extract different properties of that joint distribution. However, since skill and accuracy are multidimensional, no single measure can capture all of the dimensions. In addition, given the same set of forecasts to evaluate, meteorologists and psychologists would most likely use different measures of skill to support their study objectives. In this study, we considered approaches from both disciplines.

From meteorology, we adopted multiple measures of forecast accuracy based on the Brier Skill Score and its various components (Brier and Allen 1951; Murphy 1988, 1972a,b; Yates 1982; Stewart 1990). The measures used are described below and their properties are summarized in Table 2.

a. Brier skill score.

The Brier Skill Score is 1.0 for perfect forecasts, 0.0 if the forecast is only as accurate as the base rate (sample climatology) forecast and negative if the forecast is less
TABLE 2. Properties of skill measures.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Best possible value</th>
<th>Worst possible value</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skill score</td>
<td>1.0</td>
<td>−∞</td>
<td>0.0 is same as base rate</td>
</tr>
<tr>
<td>Squared correlation</td>
<td>1.0</td>
<td>0.0</td>
<td>“Potential” skill</td>
</tr>
<tr>
<td>Conditional bias</td>
<td>0.0</td>
<td>+∞</td>
<td>Bias in standard deviation</td>
</tr>
<tr>
<td>Unconditional bias</td>
<td>0.0</td>
<td>+∞</td>
<td>Bias in mean</td>
</tr>
</tbody>
</table>

accurate than the base rate forecast. Sample climatology for the WISE experiment was calculated from activity in the 11 nowcast areas over the nine hours used in the study.

Murphy (1988) showed that the Brier Skill Score could be decomposed into three components:

$$SS = (r_{yo})^2 - [r_{yo} - s_o/s_o]^2 - [(Y - \bar{O})/s_o]^2,$$  \hspace{1cm} (1)

where $r_{yo}$ is the correlation between forecast and observed event; $s_o$ and $s_o$ are the standard deviation of forecast probability and observed frequency of the event, respectively; $Y$ is the mean forecast probability; and $\bar{O}$ is the sample climatology of the event. Each term in the decomposition will be discussed briefly.

b. Correlation

The first term in Murphy's decomposition is the squared correlation coefficient. It is a measure of the degree of linear association between the forecast and the observed event. For probability forecasts, the correlation is computed between the forecast probabilities and the corresponding observed variable that has been assigned a value of 0.0 for nonoccurrence of the event and 1.0 for the occurrence of the event. The correlation measures the ability of the forecast to discriminate occurrence of the target event from nonoccurrence, for example, to distinguish between areas that will and will not experience severe weather during a specified period of time.

The correlation is equal to the skill score only when the second and third terms, which measure conditional and unconditional bias, are 0.0. Murphy (1988) states that since these terms are nonnegative in Eq. (1), the correlation can be considered to be a measure of “potential” skill of the forecast, that is, the skill that the forecaster could obtain by eliminating bias from the forecast.

c. Conditional bias

Murphy called the second term “conditional bias.” It might also be called “regression bias” because it indicates how well the standard deviation of the forecasts reflects the lack of perfect correlation between the forecast probabilities and observed frequency of events. A potential source of conditional bias is a tendency to make extreme probability forecasts when a predictor takes on an extreme value, even though the relation between the predictor and the variable being forecast is not strong (Hogarth 1987).

d. Unconditional bias

Murphy called the third term “unconditional bias.” For probabilistic forecasts, this term could be called the “base rate bias” because it measures how well the mean forecast matches the base rate (i.e., sample climatology). For the county nowcasts included in the limited and full information condition results presented in this paper, the base rate probability of occurrences were .32 for significant/severe convection, .22 for severe convection, and .05 for tornadoes. If the mean probability forecasts matched these base rates, then the unconditional biases would be zero. For additional numerical examples and interpretation of the skill score and its components, see Stewart (1990).

While Murphy's decomposition is a useful way of analyzing the joint distribution of forecasts and events, psychologists [such as Tucker (1964) and Stewart (1990)] have shown that consistency is also an important element of skill. When the term “consistency” is used in connection with human judgment, it refers to the extent to which identical information leads to identical judgments. A weather forecaster who was perfectly consistent would, in theory, produce identical forecasts if the same weather conditions were to occur twice (in psychology, the degree of similarity between judgments made under identical circumstances is called “reliability.” We will use the term “consistency” here because in meteorology, “reliability” has a different meaning). Since no two weather situations are ever identical, it is not possible to make direct assessments of the consistency of operational forecasts. There are techniques, however, for making indirect inferences about consistency which will be discussed in the analysis of results.

4. Results and discussion

Figures 4 and 5 show the individual skill scores and squared correlations for forecasts of significant/severe weather for the Boulder and Hanscom groups, respectively. This was the only category of weather event that was forecast in all three phases of the experiment. The designations A through E and F through L on these
scores, $F[1, 10] = 8.93, p = 0.014$; correlations, $F[1, 10] = 11.06, p = 0.008$). Local forecast experience probably played a role in this, but access to the DARE-I workstation may have also contributed to the higher scores of the Boulder forecasters in the full information condition. Comparison of Figs. 5a and 5b appears to show that the increase in squared correlation with the amount of available information is greater for the Hanscom than for the Boulder forecasters. It is important to note, however, that the difference in skill scores and squared correlations between the two groups did not change significantly as a function of the information condition (skill scores, $F[2, 20] = 0.060, p = 0.944$; squared correlations, $F[2, 20] = 0.140, p = 0.868$).

Any attempt to draw conclusions from a direct comparison of the Boulder and Hanscom groups is confounded, apart from the limited sample size, by the local forecasting experience of the Boulder contingent on the one hand, and that group's exclusive access to an advanced computer workstation in the full infor-

Fig. 4. Individual skill scores in each information condition for forecasts of significant/severe convection for (a) Boulder and (b) Hanscom forecasters.

figures refer to individual forecasters. Mean values are shown by the heavy line in each figure. Mean skill and squared correlations for each group increased with increasing information. Using the Friedman test,² the improvement was not statistically significant for either the Boulder group (skill scores, chi square $[2] = 3.60, p = 0.165$; squared correlations, chi square $[2] = 2.80, p = 0.247$), or the Hanscom group (skill scores, chi square $[2] = 4.57, p = 0.102$; squared correlations, chi square $[2] = 4.57, p = 0.102$). In fact, two forecasters in each group actually achieved greater skill in the “maps with specified cues” condition than in the full information/slide conditions.

Skill scores and squared correlations of the Boulder forecasters were significantly higher than those of the Hanscom group over all information conditions (skill

2 The Friedman test is a nonparametric test that produces a test statistic approximately distributed as a chi-square value (Siegel and Castellan 1988, pp. 174–183). In this experiment $p < 0.05$ was considered a requirement for statistical significance.

Fig. 5. Individual squared correlation scores in each information condition for forecasts of significant/severe convection for (a) Boulder and (b) Hanscom forecasters.
formation condition (as opposed to slides) on the other. Because the Boulder forecasters generally had substantially better scores in the limited information conditions (identical to those at Hanscom) for forecasts of significant/severe events, it may be suggested that a large part of their advantage was a function of their local experience. Results of tornado forecasts underscore this point. Figure 6 shows individual and squared correlation scores for tornado forecasts in the maps with specified cues and full information/slide conditions for Boulder (Fig. 6a) and Hanscom (Fig. 6b), respectively. All forecasters showed negligible potential skill using the maps with specified cues dataset. The additional information made available to Hanscom forecasters (Fig. 6b) in the form of slides made little difference in their ability to correctly forecast tornadoes. In contrast, squared correlations for all five Boulder forecasters improved in the full information condition. This result was statistically significant ($F[1, 10] = 22.59, p = 0.001$) for the group as a whole. The possibility that the Boulder forecasters also benefitted substantially from the DARE-I workstation in making the tornado forecasts cannot be ruled out. It may be that the combination of local forecast experience and ability to use the workstation as an important forecasting tool was responsible for the improvement in Boulder tornado forecast results.

As stated earlier, forecaster consistency is an important measure of skill over and above those elements contained explicitly in the decomposed Brier skill score. While it is very difficult to measure directly, it can be inferred from forecaster agreement and consensus forecasts, both of which can be easily calculated from the available data. To look at the question of interforecaster agreement, all probability forecasts for significant/severe weather were considered. In each information condition, the correlation between the probability forecasts for each possible pair of forecasters (10 pairs in Boulder and 21 pairs at Hanscom) was calculated. The highest, lowest, and median correlations for the ten pairs of Boulder forecasters are shown in Fig. 7. Mean agreement in the full information condition (mean $r = 0.68$) was less than in the limited information conditions (mean $r = 0.77$). For eight of the ten possible pairings of the five meteorologists, agreement was lowest in the full information condition. The relation between information condition and agreement was statistically significant (Friedman test, chi square $[2] = 8.60, p = 0.014$). At Hanscom (results not shown) agreement was significantly lower for the slide condition (mean $r = 0.60$) than for the limited information conditions (mean $r = 0.75$) (Friedman test, chi square $[2] = 18.67, p = 0.000$). For 18 of the 21 possible pairs of meteorologists, agreement was lowest in the full information condition.

To calculate consensus forecasts, forecast probabilities issued by each participant were standardized. The resulting forecasts were then averaged across forecasters for each of the 99 forecasts in the significant/severe category. Squared correlation scores were then computed on the basis of these consensus forecasts for each group in each information condition. The key finding is that, in the full information condition, the consensus forecasts were better than the best individual forecast for each group (Boulder consensus $r^2 = 0.286$, best individual $r^2 = 0.255$; Hanscom consensus $r^2 = 0.247$, best individual $r^2 = 0.243$). Because consensus forecasts average out inconsistencies in individual forecasts, they are more consistent than the forecasts of individ-

---

3 Standardization equalizes forecaster variance, so all forecasters receive equal weight. Because standardized forecasts can exceed 100%, they are not probabilistic. Therefore, they cannot be used to calculate skill scores and biases.

4 In most fields, this type of forecast would be referred to as "aggregate": the term "consensus" would be used only if there was genuine collaboration among the participants. We will use "consensus" in this paper because, in meteorology, it is the accepted term for describing the combination of independently made forecasts.
Fig. 7. Maximum, minimum, and median correlations of the ten possible pairs of Boulder forecasters.

Consequently, these results suggest that individual meteorologists did not reach the theoretical limit of skill and that their skill could be improved by increasing the consistency of their forecasts.

The results can be interpreted in the context of the conceptual model described in section 1. The model suggests that a) forecast skill increases with increasing information; b) there is a cognitive gap between the theoretical maximum skill and actual demonstrated skill that increases with increasing information; c) that the possible causes of the gap include increasing forecaster bias and decreasing forecaster consistency with increasing information.

Although there were some individual exceptions, forecast skill generally increased with increasing information for Boulder and Hanscom meteorologists. However, these increases were not statistically significant for either group. It seemed reasonable to expect that any gains between the map and maps with specified cues conditions would be incremental; even with storm profiles the forecasters had very little information to work with. It is somewhat surprising, however, that the full information and slide conditions did not result in significant increases in skill for the respective groups. This is particularly true in light of the fact that slides provided Hanscom meteorologists with substantially more information than they had in the limited information conditions, while the advanced workstation increased the data available to Boulder forecasters by several orders of magnitude. Another advantage of the full information/slide conditions that theoretically should have boosted scores was that forecasters could see new storm development as it was occurring and adjust their forecast probabilities accordingly (this was not the case in the limited information conditions). It might be argued that the small sample size would make it difficult to reveal a statistically significant improvement in scores even if one had occurred. However, 4 of 12 forecasters actually did better with limited rather than with full information. In addition, though the maximum percent-correct scores were best for both groups in the full information/slide conditions, the margin of improvement over a) the scores in the limited information conditions; and b) the percent correct (68%), based solely on the sample base rate of significant/severe events, was extremely small (see Table 3).

It is important to consider whether the forecasters could have been expected to do better. The fact that they may have, in fact, reached the realistic "ceiling" of skill had to be considered, given the nature of the rare phenomena they were asked to predict. However, the results from both groups strongly suggest the existence of a gap between actual and theoretical maximum forecast skill, particularly in the full information and slide conditions. The most compelling argument for this is the fact that the squared correlation scores for consensus forecasts were better than even the best individual scores in the full information and slide conditions.

Since we have no independent measure of the theoretical limits of skill for the information conditions used in this study, we are not able to provide direct evidence that the cognitive gap increased with increasing information. However, the fact that consensus forecasts produced the best results in the slide/full information condition clearly indicates that the meteorologists did not reach the limit of skill. As noted above, the primary advantage of consensus forecasting is the resulting improvement in consistency over forecasts made by individuals, suggesting that part of the gap between actual and theoretical skill is due to inconsistency. This is further supported by the fact that correlations among the forecasters in each group were lowest in the full information and slide conditions (correlations between forecasts are reduced by inconsistency).

While forecaster inconsistency was suggested as a primary cause of the cognitive gap in this study, forecaster bias did not play a major role. The lowest level of bias was found in the full information/slide condition, indicating that conditional bias improved with additional information.

For several reasons, some caution must be taken in interpreting the results of this study. The reader should be aware of the following concerns [elaborated upon in Stewart et al. (1992)], each of which will be addressed in future research.

First, differences among experts are pervasive, and in this research they increased as the information improved. The relation between information and skill was

---

5 In order to calculate percent correct scores, categorical forecasts were defined at various probability thresholds for each forecaster.

6 Using the sample base rate of significant/severe events (.32), always forecasting "no" would yield a percent correct of 68%.
TABLE 3. Maximum percent correct scores for individual forecasters in each information condition for significant/severe events.

<table>
<thead>
<tr>
<th></th>
<th>Maps</th>
<th>Maps + Cues</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boulder</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>74.75%</td>
<td>71.72%</td>
<td>75.76%</td>
</tr>
<tr>
<td>2</td>
<td>71.72%</td>
<td>72.73%</td>
<td>75.76%</td>
</tr>
<tr>
<td>3</td>
<td>70.71%</td>
<td>68.69%</td>
<td>73.86%</td>
</tr>
<tr>
<td>4</td>
<td>72.73%</td>
<td>73.74%</td>
<td>73.74%</td>
</tr>
<tr>
<td>5</td>
<td>68.69%</td>
<td>73.74%</td>
<td>75.76%</td>
</tr>
<tr>
<td>Hanscom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>68.69%</td>
<td>67.68%</td>
<td>74.75%</td>
</tr>
<tr>
<td>7</td>
<td>71.72%</td>
<td>69.70%</td>
<td>71.72%</td>
</tr>
<tr>
<td>8</td>
<td>67.68%</td>
<td>69.70%</td>
<td>70.71%</td>
</tr>
<tr>
<td>9</td>
<td>70.71%</td>
<td>72.73%</td>
<td>69.70%</td>
</tr>
<tr>
<td>10</td>
<td>70.71%</td>
<td>71.72%</td>
<td>73.74%</td>
</tr>
<tr>
<td>11</td>
<td>69.70%</td>
<td>69.70%</td>
<td>72.73%</td>
</tr>
<tr>
<td>12</td>
<td>70.71%</td>
<td>73.74%</td>
<td>74.75%</td>
</tr>
</tbody>
</table>

not the same for all meteorologists. We have checked to be sure that the mean results reflect the dominant pattern among the participants. Where individual results are not shown, we have indicated how many meteorologists fit a given result.

Second, only one forecasting problem has been studied. It is one of the most difficult problems that weather forecasters face, and their results reflected that. Even with large amounts of information, skill in forecasting severe weather probability was generally marginal.

Third, the results are based on a limited sample of forecasts. Days with interesting weather were selected for the study, but they may not be representative of all days when significant or severe weather could be a concern. Although 99 sets of forecasts were obtained, they were subject to varying degrees of temporal and spatial autocorrelation, so the effective size of the forecast set is smaller.

Fourth, the forecasting burden in the map condition (99 forecasts for one weather category) was less than for the maps with specified cues and full information conditions (99 forecasts for three weather categories, or a total of 297 forecasts). This disparity was consistent with the amount of data forecasters had to work with; participants were not asked to forecast the type of event in the map condition because they had so little information to work with.

Finally, the problem of verification is always paramount in evaluating forecasts of rare events. Though RT-87 chase teams and other spotters collectively provided excellent coverage on the two study days, it is possible that some significant and/or severe events within the study area were not detected. This would have had a bearing on the results of this experiment.

5. Concluding remarks

The results of the present study suggest that the relation between information and skill in forecasting severe weather is complex. In the context of the conceptual model depicted in Fig. 1, the implications of these results for operational forecasting depend on a) the size of the gap between actual skill and the theoretical limit of skill, and b) the cost, feasibility, and effectiveness of training and improved forecasting procedures for reducing the size of the gap. These factors will be explored in subsequent research.

Given the implications of this study, greater improvement in forecasting might be obtained by devoting resources to improving the use of information over and above those needed to increase the amount of information. Furthermore, as the amount of information increases, improving the use of that information becomes even more important.

Acknowledgments: This work was supported by the National Science Foundation under Grants SES-8822319 and SES-9109594. The participation of the 12 volunteer meteorologists who made nowcasts and the three who helped develop storm profiles and contours is gratefully acknowledged. The Forecast Systems Laboratory (FSL) of the Environmental Research Laboratories, National Oceanic and Atmospheric Administration provided weather data and workstation time. The cooperation and support of the FSL and the Atmospheric Sciences Division of the Phillips Laboratory, Hanscom Air Force Base, is also gratefully acknowledged. Special thanks to Don Chisholm of the Phillips Laboratory for his support and encouragement throughout the study, and to Gregg Hakim, for his assistance in preparing materials and pretesting a portion of the experiment. John Flueck provided suggestions during the early stages of the design of this experiment.

REFERENCES


