AD-A254 216

AWS/TR--91/001

2

AIR WEATHER SERVICE

PROBABILITY FORECASTING

A Guide for Forecasters and Staff Weather Officers

DECEMBER 1991

UNITED STATES AIR FORCE
AIR WEATHER SERVICE
Scott Air Force Base, Illinois 62225-5008

REVIEW AND APPROVAL STATEMENT

AWS/TR--91/001, *Probability Forecasting: A Guide for Forecasters and Staff Weather Officers*, December 1991, has been reviewed and is approved for publication.

ADRIAN A. RITCHIE, JR, Col, USAF
Director of Technology
Air Weather Service

FOR THE COMMANDER

WALTER S. BURGMANN
Scientific and Technical Information Program Manager
Air Weather Service

A NOTE TO OUR CUSTOMERS--

The authors and editors of this publication welcome feedback, both positive and negative. We value your opinion. Please let us know what you like and do not like about our products. Also, please let us know if your address has changed, or if you wish to receive more or fewer copies of AWS and USAFETAC technical documents in primary distribution. If you need more copies of this document, or if you know of someone else who might be interested in this or other AWS/USAFETAC publications, let us know that, too. Call, write, or FAX:

USAFETAC/LDE,
Scott AFB, IL 62225-5458

DSN 576-6648 Commercial 618 256-6648 FAX 3772

REPORT DOCUMENTATION PAGE

2. Report Date: December 1991

3. Report Type: Technical Report

4. Title: Probability Forecasting: A Guide for Forecasters and Staff Weather Officers

7. Performing Organization Name and Address: USAF Environmental Technical Applications Center (USAFETAC), Scott AFB, IL 62225-5438

8. Performing Organization Report Number: AWS/TR--91/001

9. Sponsoring/Monitoring Agency Name and Address: Air Weather Service Directorate of Technology (AWS/XT), Scott AFB, IL 62225-5008

10. Sponsoring/Monitoring Agency Report Number: AWS/TR--91/001

11. Supplementary Notes: Supersedes AWS Pamphlet 105-51, 31 October 1978. Includes Change 1, 16 October 1989.

12a. Distribution/Availability Statement: Approved for public release; distribution is unlimited.

13. Abstract: Describes recommended techniques for producing and evaluating probability forecasts. Includes a selected number of applications for optimal decision-making. Chapters 1-4 address specific needs of forecasters and supervisors who make and evaluate subjective probability forecasts. Chapters 5-7 address applications of probabilities in decision-making, primarily for staff weather officers and staffmets.

14. Subject Terms: *METEOROLOGY, *WEATHER, WEATHER FORECASTING, PROBABILITY FORECASTING, CLIMATOLOGY, BRIER PROBABILITY SCORE

15. Number of Pages: 124

17. Security Classification of Report: UNCLASSIFIED

18. Security Classification of this Page: UNCLASSIFIED

19. Security Classification of Abstract: UNCLASSIFIED

20. Limitation of Abstract: UL

Standard Form 298

PREFACE

With the reorganization and restructuring of the Air Weather Service in 1991, AWS Pamphlet 105-51 was selected for conversion to AWS Technical Report 91/001. The conversion was intended to preserve the document when the AWS Pamphlet is rescinded, as well as to make it universally available through the Defense Technical Information Center (DTIC).

DEPARTMENT OF THE AIR FORCE
Headquarters Air Weather Service (MAC)
Scott Air Force Base, Illinois 62225

Formerly AWS PAMPHLET 105-51

31 October 1978

Weather

PROBABILITY FORECASTING: A Guide for Forecasters and Staff Weather Officers

This pamphlet describes recommended techniques for producing and evaluating probability forecasts. It also includes a selected number of applications for optimal decision making.

	Paragraph	*Page*
Chapter 1—Introduction		
General	1-1	1-1
References	1-2	1-1
Terms/Definitions	1-3	1-1
Types of Meteorological Probabilities	1-4	1-1
Chapter 2—Why Probability Forecasts?		
General	2-1	2-1
Characteristics of Categorical Forecasts	2-2	2-1
Characteristics of Probability Forecasts	2-3	2-1
Reasons for Adoption	2-4	2-2
Chapter 3—How to Prepare Probability Forecasts		
General	3-1	3-1
Defining the Event	3-2	3-1
Precision of Probability Forecasts	3-3	3-1
Preparing the Forecast	3-4	3-1
Amending Probability Forecasts	3-5	3-5
Chapter 4—Evaluation Techniques		
General	4-1	4-1
Sharpness and Reliability	4-2	4-1
Controlling Sharpness and Reliability	4-3	4-5
Brier Probability Score (PS)	4-4	4-12
Summary	4-5	4-26
Chapter 5—Probabilities in Decision Making		
Introduction	5-1	5-1
General Decision Matrix	5-2	5-1
Utilities	5-3	5-4
Original Cost-Loss Model	5-4	5-5
General Cost-Loss Model	5-5	5-6
Critical Probability	5-6	5-7
Value Analysis	5-7	5-12
Other Models	5-8	5-15
Chapter 6—Introduction to Weather Impact and Mission Success Indicators		
Introduction	6-1	6-1
Forms of WII	6-2	6-1
Weather Effect Models	6-3	6-6
Mission Success Indicators	6-4	6-11
Categories of WII	6-5	6-19
Chapter 7—Implementation of Probability Forecasts		
General	7-1	7-1
Development	7-2	7-1
Testing	7-3	7-1
Evaluation	7-4	7-1
Operational Use	7-5	7-1

OPR: DN
DISTRIBUTION: F; X (See signature page.)

Page

Attachment

1. **Bibliography** A1-1
2. **Terms Explained** A2-1
3. **Selecting Probability Intervals** A3-1
4. **Explanation of Mathematical Symbology in the Brier Score Equation** A4-1
5. **Table of Partial Brier Scores** A5-1
6. **Determining Utilities in Terms of Regret** A6-1
7. **Procedures Used in Preparing Table 5-15** A7-1
8. **Introductory Training Scenario** A8-1
9. **Introduction to Decision Trees** A9-1
10. **Instructions for Variable-Width Interval Forecasting of Maximum and Minimum Temperature** A10-1

Chapter 1

INTRODUCTION

1-1. General. This pamphlet provides the standard tools and techniques on probability forecasting. It is the basic reference for self-study, and the primary source for developing local training programs.

a. Chapters 1-4 meet the specific needs of forecasters and supervisors making and evaluating subjective probability forecasts. Commanders and staff members will find the information useful as a comprehensive reference.

b. Chapters 5-7 address applications of probabilities in decision making, and are designed primarily for staff weather officers (SWOs) and staff meteorologists. These chapters describe the more complex aspects of decision theory and weather impact indicators. Customers must have a good understanding of the advantages of probability forecasts and how they can enhance decision making, before specialized applications are attempted.

1-2. References. Some prior knowledge of probability theory and related mathematics is required to understand the first four chapters of this volume. The sections are arranged so that basic information is presented first, to aid the transition into more technical discussions. Use AWSTR 77-267, Guide for Applied Climatology, if more mathematical background is required. The references, identified by an asterisk in the bibliography (Attachment 1), are recommended for every forecasting unit.

1-3. Terms/Definitions. Basic terms and definitions are in Attachment 2. Review them now and use them as references while reading the remainder of this pamphlet. Do not be concerned initially about acquiring a total understanding of all the terms; their meaning will be clearer when they are seen in context later in the pamphlet.

1-4. Types of Meteorological Probabilities. Three types of probabilities are commonly used in meteorolgy climatological, objective, and subjective. Each type, with typical applications, is described below. Note that the term "probability forecast" in subsequent chapters refers to the subjective forecast, unless otherwise noted. However, many of the applications and evaluation techniques discussed apply to all three types.

a. Climatological Probability. The probability that an event will occur based on historical observations or experimental data. AWSTR 77-267, Guide for Applied Climatology, describes the most common methods of obtaining climatological probabilities. Many of these techniques can be applied directly at the unit level. Others, which are more complicated or need extensive data or data processing facilities require squadron, wing, or USAFETAC assistance to apply. The Revised Uniform Summary of Surface Weather Observations (RUSSWO) for each base contains many climatological probabilities.

Climatological probabilities are used primarily in planning and design functions. They are also extremely important as inputs to all forecasts, categorical and probabilistic. Examples of some planning problems that can be resolved by using climatological probabilities follow.

Example 1. What is the probability of <1000 feet ceiling and/or < 2 miles visibility at Base A in January? **NOTE:** The RUSSWO gives the climatic frequency for equal to or greater than those conditions.

Example 2. Base B can expect 12 days with 0.01 inch or more precipitation during March. What is the probability of having no more than 6, 8, 10, 12, 14, 16 days with 0.01 inch or more precipitation?

Example 3. What is the probability that Base C will be above alternate minimums (ceiling ≥1000 feet and visibility ≥ 3 miles), given that Base D is below GCA minimums (ceiling <200 feet and/or visibility <1/8 mile)?

Example 4. What is the probability that either Base A or Base B, or both, will be above alternate minimums, given that Base C is below minimums?

Example 5. An attack on a coastal installation is being planned. Troops and equipment can be delivered to the area by air, sea, or both. Given the critical weather thresholds for each, what is the probability of success considering weather constraints only? What is the probability of success by sea, given that air delivery is unfavorable?

b. Objective probability. The probability that an event will occur using a fixed set of rules which produce a unique and reproducible outcome. These rules are derived by empirical or theoretical means, or a combination of both. Objective probability techniques are assuming increased importance in operational forecasting. Three methods used by the National Weather Service in automating their terminal forecasts illustrate these techniques (Bocchieri, Crisci, et al, 1974).

Example 1. Single-station equations were developed to predict the probability of maximum, and minimum temperatures, surface wind, and cloud cover using only the weather observations at the local terminal. More than 30 possible predictors were screened and the best predictors combined into objective prediction equations.

Example 2. Model Output Statistics (MOS) uses statistical methods to complement the output of numerical models. The technique matches observations of local weather with output from numerical models. Since numerical models do not directly predict the elements of more interest to a forecaster, the MOS technique, in effect, determines the "weather related" statistics of the numerical model. For instance, it could give the probability of precipitation at a station for a corresponding model prediction of 80% relative humidity; or surface winds corresponding to a model prediction of the 1000 mb geostrophic wind. Resultant forecast equations are derived by statistical techniques. In this way, the bias and inaccuracy of the numerical model, as well as local climatology, can be incorporated into the forecast system. MOS products are produced by the NWS Techniques Development Laboratory (TDL) and include forecasts of precipitation, temperature, wind, clouds, ceiling, visibility, and thunderstorms.

Example 3. The third approach combines the output from single-station equations, forecast output from numerical models, and the MOS technique to form predictors for another set of prediction equations. These equations produce objective probability forecasts of various weather elements which are equal to or better than man-made forecasts in many instances, depending upon the element and forecast period. Elements that have been successfully forecast include maximum and minimum temperatures, surface wind, cloud cover, precipitation, ceiling, visibility, thunderstorms, and freezing precipitation.

c. Subjective probability is the personal estimate of the probability that a given event will occur. Unlike climatological and objective probabilities, there are no firm rules or techniques used in deriving a subjective probability forecast. In practice, forecasters study the available data as they would in preparing a conventional forecast, and then subjectively assign a probability value which reflects their confidence that the event will occur.

Chapter 2

WHY PROBABILITY FORECASTS?

2-1. General. This chapter discusses categorical and probability forecasts and shows how probability forecasts can enhance decision making.

2-2. Characteristics of Categorical Forecasts. A categorical forecast specifies that a given weather event will occur. The forecast can be for either a two category event (e.g., rain or no rain), or for a multicategory event (e.g., visibility 0 to 1/2 mile, 1/2 to 2 miles, 2 to 3 miles, or greater than 3 miles). Categorical forecasts cause several problems.

a. Unquantified Uncertainty. At times, forecasters are certain that an event will occur. More often, they are not. A forecaster making categorical forecasts cannot mention other possible outcomes, or express the degree of uncertainty in the forecast. Uncertainty exists for several reasons.

(1) We cannot accurately describe the initial state of the atmosphere. Observations are not available for vast ocean and land areas. Our fixed observational network provides a limited measurement, in time and space, of many (but not all) weather variables. Surface observations for specific points are not necessarily representative of large areas, or of points between reporting stations. The same is true of upper air soundings. In addition, these measurements are ascribed to the launch point even though the instrument package might be many miles away as it rises. Finally, the instruments used to measure atmospheric variables have inherent inaccuracies.

(2) The output from our dynamic prediction models is not perfect. These models often neglect potentially significant atmospheric processes. This is partially due to our imperfect knowledge of the physical processes involved and how to model them. At times, it results from our computers not being large or fast enough to incorporate these complex processes into our models.

(3) Even if atmospheric observations and computer models were accurate, it is doubtful that forecasters could always interpret these correctly and consider local modifying effects to make perfect area or point forecasts.

b. Limited Use in Decision Making. Categorical forecasts are generally made for the event that is most likely to occur (i.e., the category with the highest probability). However, there are times when the possible occurrence of certain unfavorable weather conditions is important to the customer, such as damaging hail or strong winds. For these situations, forecasters tend to intuitively use a much lower probability of occurrence threshold (for example, 10%) to differentiate between a yes/no categorical forecast. This threshold is usually based on the forecaster's estimate of the impact of the weather event on the customer's mission. Once the forecaster determines that the probability of occurrence exceeds his threshold, a categorical forecast is made which implies certainty that the event will occur. Thus, the forecaster assumes the role of decision maker. The disadvantage is that the forecaster does not have sufficient knowledge of all the operational factors that should be considered in establishing the proper probability threshold for making the decision. However, since certainty is implied, the customer should take action. In actual practice, after categorical forecasts have been issued, but before very important decisions are made, a dialogue takes place between the forecaster and decision maker. The decision maker tries to find out how confident the forecaster really is about the chances (probability) of the event actually occurring. The preceding is an example of subjective decision making. It is time consuming, requires that each case be handled individually, has no set rules, and may not produce the best decision. Categorical forecasts do not enhance subjective decision making.

Objective decision making uses a set of rules or a decision model to arrive at a decision. Given the same initial conditions, the objective decision making process will produce the same decision every time. Numerous studies show that when categorical forecasts are used in objective decision models, the long term benefit is less than when decisions are based on probability forecasts. (Murphy, 1977).

2-3. Characteristics of Probability Forecasts. Probability forecasts reflect the forecaster's perception of the state-of-the-art for predicting a particular event, given existing conditions.

a. Quantified Uncertainty. Probability forecasts quantify uncertainty. They do not eliminate the causes of uncertainty described in paragraph 2-2a; rather they allow the forecaster to express all outcomes quantitatively in probabilistic terms.

b. Optimum Use in Decision Making. Probability forecasting does not change the skill or accuracy of the forecasts, but by providing a quantitative assessment of all possibilities, does enhance decision making. Further, the forecaster concentrates on what he does best, forecasting the weather, leaving operational thresholds and decisions to the customer. The following examples illustrate typical applications of probability forecasts in various types of decisions.

Example 1. A mission scheduled for base A can use either base B or base C as an alternate. The categorical forecasts for bases B and C are for above minimum conditions. However, the base B forecast is for 55% probability of above minimums, while the base C forecast is for 90% probability. By considering the probability forecast, the decisionmaker can make a better choice of an alternate if weather is the only factor.

Example 2. During the first part of a training period, a wing commander may use a 60% probability of favorable weather as the threshold to make "go" decisions for flying training missions. Toward the end of the period, however, the commander might change the threshold probability to 40%, if training is behind schedule, or to 70% if ahead of schedule.

Example 3. A C-130 wing commander must protect base aircraft from winds greater than 35 knots. A forecaster using categorical forecasts probably will not issue a wind warning unless the probability of occurrence is higher than the probability of nonoccurrence, e.g., greater than 50%. However, the wing commander determines that the costs to protect are small compared with the possible loss, and that warnings are needed more often than this. Protective action will be taken if the probability of occurrence is greater than 30%.

Example 4. The same C-130 wing commander decides that, with a C-5 on his base, to lower the probability (and thus the risk) above which to take protective action to 10%.

Example 5. Given a 50% probability of favorable aerial refueling weather for an overseas training deployment of fighters Tactical Air Command would most likely delay the mission, or look for a refueling area with a higher probability of favorable weather. In the event of a contingency, however, a threshold probability as low as 20% may trigger a "go" decision.

c. Multiple Use. Probability forecasts allow more than one customer to use the same forecast. Customers on the same base have widely varying priorities, mission urgencies, flying experience, and aircraft with different weather sensitivities, instrumentation, and ordnance. With probability forecasts, they can weigh these factors individually and act only when the forecast probability exceeds their critical probability threshold. Consider the following examples:

Example 1. An aero club might take protective action when the probability of 30-knot winds exceeds 20%, but an F-4 wing might wait until the probability exceeds 60%.

Example 2. The forecast for a base may be for 60% probability of below landing minimums. An HC-130 on a rescue mission to this base would probably "go." A student pilot planning a cross country solo in a T-37 certainly would not plan to land at this base.

d. Problems. Although probability forecasting offers advantages, there are several potential problems with implementing this program.

(1) When the National Weather Service (NWS) started using probabilities in precipitation forecasts, they encountered three main problems: forecaster tendency to suppress uncertainty, customer lack of understanding of what probabilty forecasts actually mean, and objections to increased user/decisionmaker workload (Kelly, 1976). Similar problems will undoubtedly affect AWS efforts.

(2) Any new procedure causes an initial surge in workload to train forecasters. New educational programs must be devised. Probability forecasts require the forecaster to consider all possible weather outcomes and quantify the probability of occurrence of each. Verification of probability forecasts also requires more time and effort than verification of categorical forecasts. This increased workload need not be very large with proper training. Its extent depends on how the forecasts are implemented. In some cases, a number of customers or a variety of requirements can be satisfied by one forecast, with only a small increase in workload. The wide use of tailored probability forecasts could result in a substantial increase in workload.

(3) A major problem is customer acceptance of probability forecasts. Air Force decisionmakers are generally concerned only with their next decision; the quality of yesterday's or tomorrow's forecast does not concern them today. The key to solving this problem is convincing the customer of the benefits derived from using probability forecasts (Chapter 5). However, the fact that probability forecasts save money "in the long run" may not sway some Air Force decisionmakers to accept probability forecasts for all missions.

2-4. Reasons for Adoption.

a. Enhanced Use of Forecasting Services. If decisionmakers had perfect categorical forecasts, their decisions would be simple: select the course of action which produces the best result. It is generally conceded that we will not be able to predict weather events with perfection in the foreseeable future. Further improvements in accuracy will come in small increments, as we refine existing techniques. Therefore, we must look for better ways to enhance the use of our existing prediction capability in the customer's decision making process. This is especially important, since our weapons systems and tactics are becoming more weather sensitive, and the decision processes more complicated.

b. Potential Cost Savings. Although use of probability forecasts will not increase our forecasting skill, their increased utility for decisionmaking can lead to substantial resource savings.

Example 1. The Space and Missile Test Center (SAMTEC) manages the Western Test Range, which extends from the launch site at Vandenberg AFB, California to the Indian Ocean. The weather is extemely important when R&D ballistic missile launches are planned, because of uprange, midrange, and downrange weather constraints. Activation of all facilities and sensors necessary to support such a complex launch must begin several hours before scheduled launch time. If the operation is scrubbed late in the count-down, thousands of dollars (in some cases hundreds of thousands) in range costs are expended with no payoff. To avoid these costly "weather scrubs," SAMTEC began using probability forecasts for decisions to activate the range and continue a count-down. The probability forecasts were for specialized weather criteria that was so climatologically rare that it seldom, if ever, would have been forecast had categorical forecasts been used. By using probability forecasts, SAMTEC was alerted to those cases when the probability of occurrence was significantly higher than the climatological probability. Over a period of 14 months, SAMTEC documented a net savings of $3,200,000 in range support costs by avoiding 18 unsuccessful count-downs. (Lyon and LeBlanc, 1976).

Example 2. A study of the United States construction industry by Russo (1966) estimated that the annual dollar loss to the construction industry due to weather causes ranged from $3 to $10 billion. Using techniques similar to those described in Chapter 5, Russo determined that an annual savings of $0.5 to $10 billion was possible, if probabilistic forecasts of critical weather elements were provided to and used appropriately by the industry. Skill levels existing at that time were assumed. Russo also found that the maximum achievable savings, assuming 100% accuracy of all short range forecasts (0-24 hours), was only $300 million above that of probabilistic forecasts.

These examples illustrate how significant savings are obtained by using probability forecasts in weather sensitive decisions. Since weather affects almost every facet of military operations, there is no reason why similar savings cannot be achieved in this area as well.

Chapter 3

HOW TO PREPARE PROBABILITY FORECASTS

3-1. General. The meteorological principles used to prepare categorical forecasts also apply to probability forecasting. Any forecaster capable of producing good categorical forecasts can also produce good probability forecasts by following a few simple guidelines. This chapter describes how to prepare probability forecasts, and offers suggestions for amending them.

3-2. Defining the Event. The forecast event must be precisely defined and understood by both the customer and the forecaster. The importance of this must not be underestimated. Users will assign a variety of interpretations to a single probability forecast if the event is not precisely identified. Myers (1974) listed a total of six different interpretations of the meaning given to probability of precipitation (POP) forecasts by the public.

(1) The probability that measurable rain (i.e., 0.01 inch or more) will fall somewhere within the forecast area sometime during the period covered by the forecast:

(2) The probability that a general rain will cover the entire area;

(3) The fraction of the forecast area that will receive measurable rain in the forecast period;

(4) The fraction of the time interval during which measurable rain falls;

(5) The probability that a traveler in the forecast area will encounter rain during the forecast period; and

(6) The probability that a specific point in the forecast area will receive measurable rain sometime during the forecast period. This is the official definition, but even it is not clearly understood or used by all forecasters (Murphy and Winkler, 1974).

a. Tailoring Forecasts. Operations require forecasts tailored to specific requirements. This means the event must be defined in terms of a weather element exceeding a certain threshold (amount, duration, intensity, etc.). For example, the Base Civil Engineer may require predictions of the most probable rainfall amount, the number of hours during which a given intensity of rainfall will occur, or the probability of total rainfall exceeding a specified amount. To another customer a 15% chance of freezing rain may be more significant than an accompanying 70% chance of light rain and 5% chance of sleet, all in a situation where the total probability of precipitation is 90%. The important point is that the event must be stated in terms of the likelihood of the element exceeding a critical threshold.

b. Determination of Forecast Periods. The time period is an important factor to consider when preparing a probability forecast. For many cases the forecaster will be confident that an event will occur, but will be uncertain about the actual timing. Consider the following example where a cold front with a well defined rain band is approaching a base. The event to be forecast is the occurrence of rain at the base any time during a six hour forecast period.

The forecaster believes that there is a 100% probability that rain will occur at the station and will last only one or two hours. He is uncertain, however, exactly when it will occur. If the time of occurrence is centered around the dividing time between forecast periods, three possibilities exist: (1) all the rain may fall during the first forecast period, (2) all of it may fall during the second forecast period, or (3) it may rain during both periods. In addition, if the midpoint of the rain period is expected to be exactly on the dividing time between forecasts periods, each of the three possibilities is equally likely. Thus, there are two out of three chances (67% probability) that it will rain in the first period with the same probability for the second period. Thus, the 100% probability of occurrence becomes 67% for each of the fixed time periods (Hughes, 1965).

If we change the event to rain at the 6th hour of a forecast period, the same three possibilities exist. In this case, however, the probability of occurrence becomes 33% (one chance in three of rain occurring at the 6th hour).

Conversely, if the forecaster is confident about the timing of the event, and the duration is expected to be much less than the forecast period, it would be best to assign various probabilities to increments of the forecast period. For example, the probability of precipitation for an eight hour period may be 60%, but the probabilities for two hour increments of the forecast period could be 50%, 30%, 20% and 10%, respectively. Note that if the probability forecasts are made for increments of the forecast period, the sum of the probabilities may exceed a single probability forecast for the entire forecast period, and may even exceed 100%, since the events in this case are not mutually exclusive.

3-3. Precision of Probability Forecasts. Any probability value from 0-100% can be used for forecasting purposes, but the use of all integers between 0 and 100 implies more precision than actually exists. The forecast increments should be as detailed as required by the customer, but should not be more precise than is justified by forecasting skill. Except for values near the extremes, forecasters generally cannot differentiate much finer than 10% probability increments. However, for rare events, probability increments must be small enough to allow forecasters to select probability values on both sides of the climatic frequency of the event. The size of the probability increments will also affect forecast verification, since for verification purposes it is desirable to group probability forecasts into intervals which correspond to the probability increments that will be used. For information about how NWS selects probability intervals, see Attachment 3.

3-4. Preparing the Forecast. The process of analyzing meteorological data is essentially the same when preparing either categorical or probability forecasts. When preparing a categorical forecast, the forecaster must predict the conditions most likely to occur during the forecast period. However, for a probability forecast, he quantifies the likelihood of a specific, predefined event occuring during the forecast period.

The forecaster must consider such factors as the climatic frequency of the event, the size of the forecast area, and the expected timing and duration of the event. When assessing the probabilities, the forecaster must think in terms of groups of forecast situations and compare the present meteorological conditions to those experienced in the past.

For example, if a forecaster knows that a given synoptic situation produced rain every time it occurred in the past, and that the exact condition exists today, then the forecast probability should be 100%. On the other hand, another meteorological situation may have produced rain on 6 out of 10 times in the forecaster's past experience. If similar conditions exist today, the probability should be 60%.

a. Use of Long-Term Climatology. Climatology is the starting point for every probability forecast. Over the long term, the weighted average of the forecast probabilities should equal the climatic probability of the event (assuming no climatic change and that the forecasts are reliable). A desired objective of probability forecasting is to move individual probabilities away from climatology. Climatic probabilities tell the forecaster how frequently high and low probabilities should be used (i.e., sharpness distribution). Consider the RUSSWO climatology for Scott AFB given in Table 3-1.

Table 3-1. Climatic Probabilities for TAF Ceiling Categories for Scott AFB. Valid 1800Z Dec.

CATEGORY	A	B	C	D
PROBABILITY	.00	.10	.21	.69

The climatic probabilities imply that most forecasts for category A should be for probability values near zero. Similar reasoning applies for category B. However, the frequency of high forecast probability values would be quite large for category D. If the forecasts for category D were perfect, there would be 69 forecasts out of 100 with a probability of 100%, and 31 out of 100 with a probability of 0%. However, it is unrealistic to expect such sharpness in most cases.

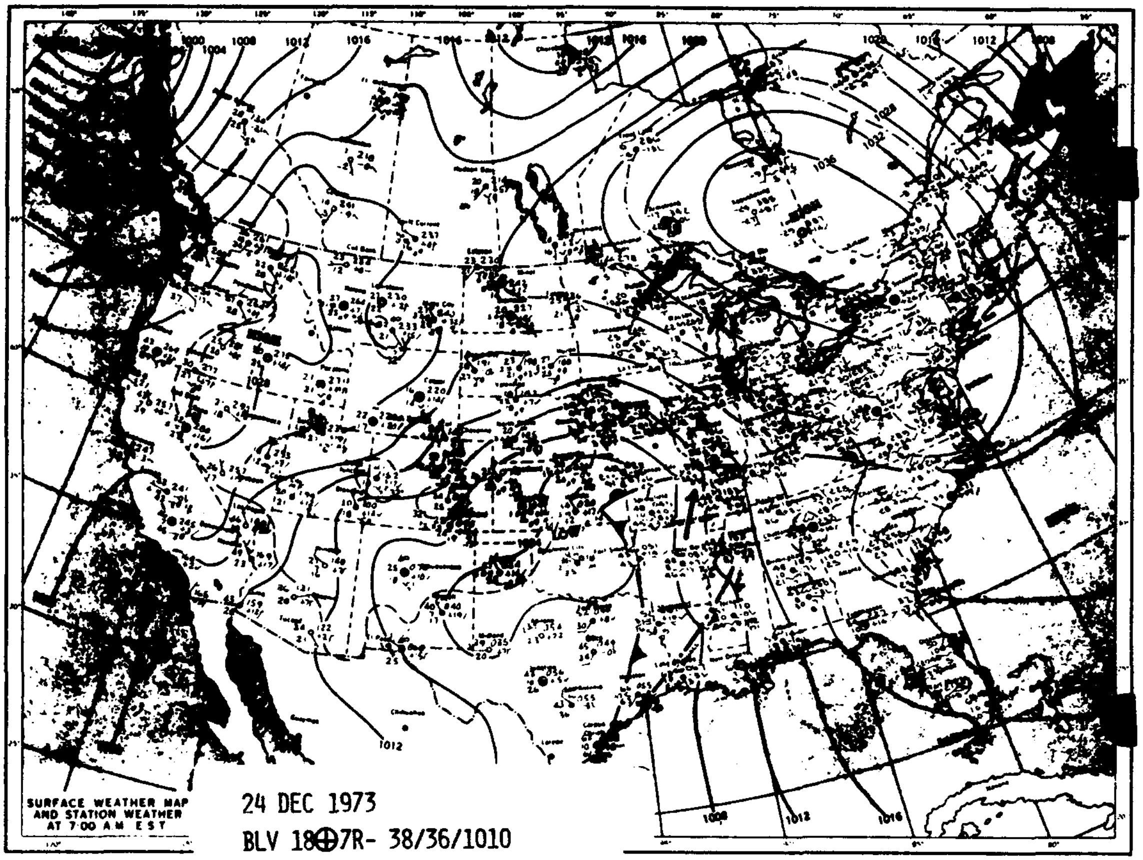
b. Use of Conditional Climatology (CC). For ceiling and visibility forecasts, most units have CC tables which provide a starting point with built-in skill. It is a challenge for most forecasters to surpass the forecasting skill of these tables. There are several kinds of CC tables (unstratified, stratified, etc.), but there is no one best kind for all situations.

(1) Conversion of CC categories to TAF categories. One minor difficulty in using the older CC tables is that the categories are not the same as those presently used in TAFs. Table 3-2 shows how to convert the probabilities in older CC tables to existing TAF categories.

Table 3-2. Six Hour CC Conversion Table

ENTER WITH:				OBTAIN:	
				EQUIVALENT TAF CATEGORY AND CC PROBABILITY	
INITIAL CIG CC CATEGORY	CC CAT	CC PROB	CIG/VSBY LIMITS	TAF CAT	PROB
A	A	13	<200	A	13
	B	13	≥200<500	B	40
	C	27	≥500<1000		
	D	27	≥1000<3000	C	27
	E	7	≥3000<10,000	D	20
	F	13	≥10,000		
INITIAL VSBY CC CATEGORY					
J	J	13	<½	A	13
	K	8	≥½ <1	B	32
	L	24	≥1 <2		
	M	5	≥2 <3	C	5
	N	18	≥3 <6	D	50
	O	32	≥6		

(2) Example Using CC. The following example shows how CC tables are used to prepare probability forecasts. Consider a six hour forecast of the four ceiling categories in the TAF for Scott AFB. The forecast will be made by using the 0700 EST surface charts (Figure 3-1) and will be valid for 1300 EST on 25 December. The surface chart for the previous day is provided for continuity. Observations at map time are written at the bottom of the charts. Arrows on the charts point toward plotted observations for St Louis MO. The long term climatic probabilities for the TAF categories are: A - 0%, B - 10%, C - 21%, and D - 69%. Wind stratified CC probabilities for this situation are as follows: A - 13%, B - 40%, C - 27% and D - 20% (Note that the occurrence of TAF categories are mutually exclusive events, so the sum of the probabilities for TAF categories always equal 100%.) CC probabilities make a reasonably good forecast. The forecaster must determine how much (if any) the CC probabilities must be adjusted for the particular situation. In this case CC indicates that the probability of the initial category (A) remaining for six hours is only 13%, and the most likely category to occur is B. But since a cold front over Scott AFB is not an average situation, and continuity suggests a clearing trend after the frontal passage, one might expect the CC values to be on the pessimistic side. Timing of the frontal passage in this case is the major uncertainty. Rather than assigning a probability of 100% to Category D, the timing uncertainty can be accounted for by adjusting the CC probabilities as follows: A - 0%, B - 5%, C - 15%, and D - 80%. Other forecasters may have chosen different values based on their experience and confidence. Category D verified. In this example CC indicated the trend, but since the clearing was caused by a relatively unusual situation, CC was pessimistic and overforecast categories B and C. CC probabilities must be modified when the existing situation is not average. Even then there should be a good reason for deviating. This does not imply that the well-known biases of CC, e.g., weakness in forecasting downtrends, should be ignored. In summary, use CC tables as a starting point for distributing probabilities, when more than two categories are involved.



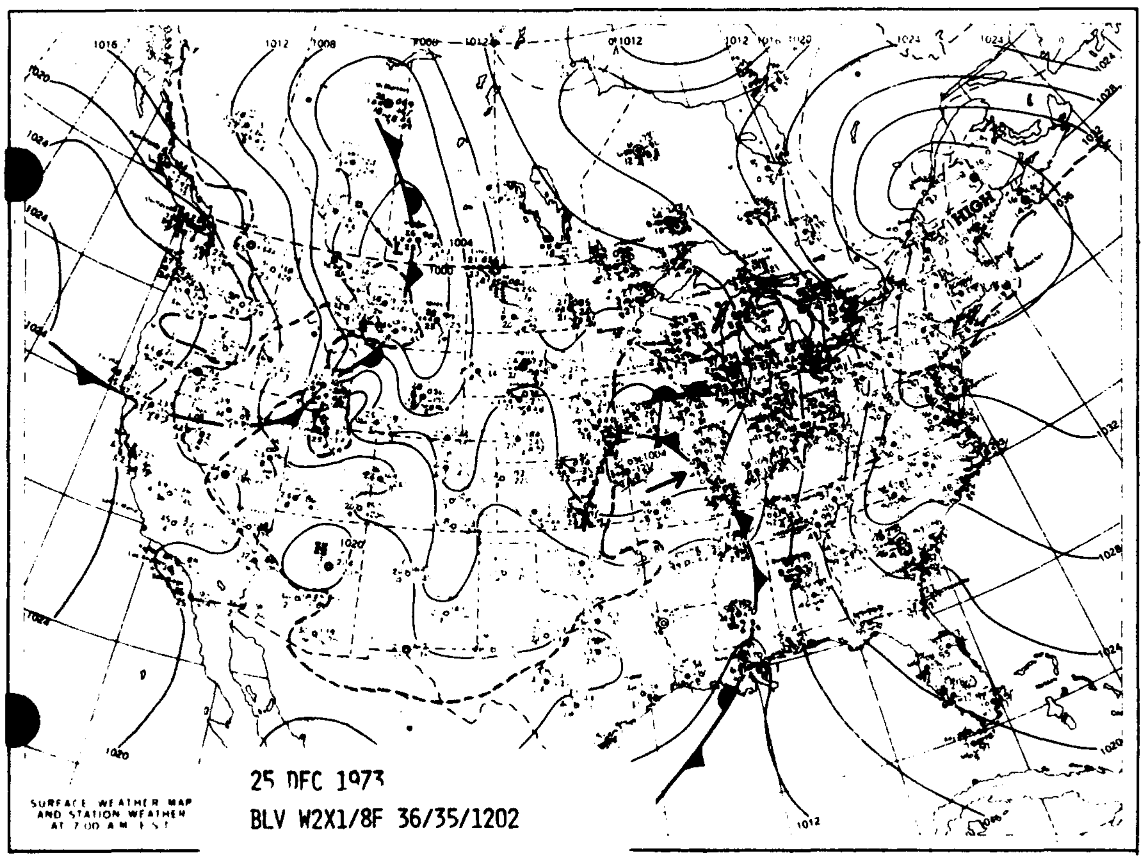


Figure 3-1. Surface charts for preparing a forecast using conditional climatology.

c. Use of Objective Forecast Studies. Many local forecast studies contain guidance already stated in probabilistic terms (observed frequency). Other studies may be converted for use in probability forecasting. The utility of these aids can be evaluated using techniques described in Chapter 4.

d. Centralized Forecaster Aids. Centrally produced probability forecasts, such as TDL MOS forecasts are a valuable input to local forecasts, if adjustments are made to account for known model biases and recent observations. Rules for modifying objective forecasts may be developed, but modifications should not be made unless there is good reason to do so. The centralized forecasts are especially useful beyond the 12 hour point. Experiences of the Central Region of NWS indicate that their forecasters can successfully improve upon portions of TDL MOS forecasts, but most improvement occurs only during the first 12 hours.

3-5. Amending Probability Forecasts. Confidence that the event will or will not occur increases as the lead time in a probability forecast erodes. This change of confidence means that amendment procedures must be established. The first approach should be to avoid amendment problems by issuing updates at prescribed times. If this cannot be done, then establish rules which specify amendment criteria. User requirements and the type of forecast will determine the amendment criteria for each case. The following amendment criteria might apply.

a. When it appears that a TAF category other than the one with the highest probability will verify.

b. When the forecast probability passes through the customer's critical probability in either direction.

c. When the forecast probability changes by a specified interval, for example if ±20%.

Chapter 4

EVALUATION TECHNIQUES

4-1. General. The techniques for evaluating probability forecasts are different and more complicated than for categorical forecasts. However, the objectives are the same: to determine how good the forecasts are and to show how to improve them. Verification feedback to those who prepare probability forecasts is a key element in the evaluation process. It is also important for the decision maker who receives probability forecasts to review verification data periodically, since the quality of the forecasts affects his thought process. This chapter describes techniques for evaluating probability forecasts and how to improve them. Sharpness and reliability, two properties of probability forecasts, are discussed. Methods for measuring and achieving good sharpness and reliability are shown. The chapter concludes with a discussion of the Brier probability score, a system for computing a single number that reflects the overall goodness (sharpness and reliability) of a set of probability forecasts. While reading the chapter, keep in mind that the purpose is not to impose all of the verification schemes shown, but to show the methods that could be employed.

4-2. Sharpness and Reliability. In order to evaluate a set of probability forecasts, one must consider two properties: sharpness and reliability. Sharpness is the ability to "sort" all possible events into an ordered set of categories of likelihood of occurrence (e.g., rain or no rain) (Sanders, 1963). Resolution is another term sometimes used, but we prefer sharpness. Reliability is the ability to "label" each category derived in the sorting process with a specific likelihood, or probability of occurrence (Sanders, 1963). For example, the probability of rain is 65% (no rain - 35%).

a. Sharpness. Sharpness measures the degree of certainty of probability forecasts. "Perfect" sharpness occurs when all forecasts are for either 0% or 100% probability of an event occurring. Categorical forecasts have maximum certainty and, thus, have "perfect" sharpness (categorical forecasts are a special case of probability forecasts). "Zero" sharpness exists when all forecasts are for the climatological probability of the event. This is because the climatological probability (or frequency of occurrence) is generally known, and a forecaster with minimum certainty can always forecast climatology. The objective of measuring sharpness, therefore, is to determine a forecaster's ability to move the predicted probabilities away from the event's climatological frequency. It is important to note that the measure of sharpness has nothing to do with the actual occurrences of the event.

(1) Sharpness Diagrams. To measure sharpness, determine how forecasts are distributed throughout the range of probabilities (0-100%) with respect to the climatological frequency. One method is to depict on a forecast distribution graph the number of times each probability was used in the set of forecasts being evaluated. Plotting the counts in the appropriate probability interval results in a bar graph (Figure 4-1).

FORECAST VERIFICATION

Forecast Probability	OCCURRENCES	# OF FCSTS	OBSVD FREQ
100	5	7	71%
80	2	4	50%
60	1	4	25%
40	1	1	100%
20	1	6	17%
0	0	9	0%
	Totals		Climo
	10	31	32%

FORECAST DISTRIBUTION

Forecast Probability	1	2	3	4	5	6	7	8	9	10	11	12
100	1	0	1	1	1	1	0					
80	1	1	0	0								
60	1	0	0	0								
40	1											
20	0	0	1	0	0	0						
0	0	0	0	0	0	0	0	0	0			

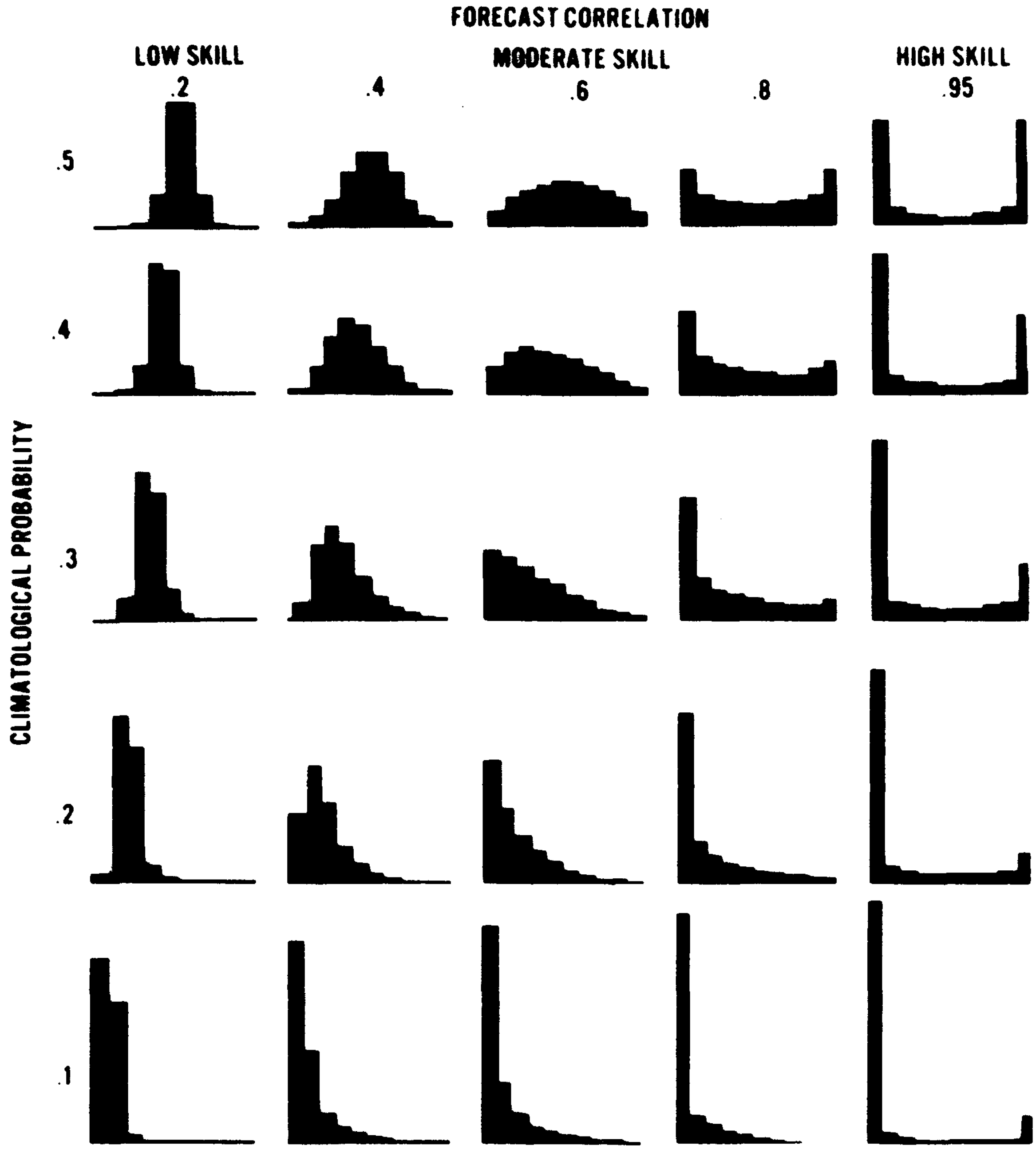
Forecast Frequency

Legend: 1-Represents event occurrences
0-Represents nonoccurrences

Figure 4-1. Example Forecast Distribution Diagram (Sharpness).

In this example, probability intervals of 20% were used; for most operational forecasts, smaller intervals are usually required. For verification, an observed event is annotated by a "1" (an observed probability of 100%); an event that did not occur is labeled with a "0" (an observed probability of 0%) (Sanders, 1958). Forecast frequency is the number of times each probability value was forecast. After the number of forecasts in each interval is plotted, bars can be drawn to highlight the distribution. If the set of forecasts is very large, one can compute the percentage of forecasts in each probability interval and plot these percentages proportionally for the forecast frequency. This diagram illustrates a sharpness pattern one might obtain from an evaluation of a series of 31 forecasts issued once daily for the occurrence of a ceiling and/or visibility below 3000 feet and/or 3 miles six hours later. This set of forecasts exhibits a fairly good degree of sharpness, i.e., 16 of 31 forecasts were in either the 0 or 100% probability intervals, with another 10 in adjacent intervals (4 in 80% and 6 in 20%). Note that only one forecast was in the interval (40%) closest to sample climatology (32%), i.e., zero sharpness was not a major problem. If these forecasts had exhibited perfect sharpness, all would have fallen in either the 0 or 100% intervals. Additionally, if the forecasts were all perfectly accurate, the forecast probabilities would have been distributed in those two intervals in proportion to the number of observed and not observed cases, i.e., all 10 event occurrences would have been in the 100% interval, and all 21 nonoccurrences in the 0% interval. This is exactly what categorical "yes or no" forecasts attempt to do. In fact, this and other discussions that follow indicate that categorical forecasts are simply a special case of probabilty forecasts.

(2) Typical Forecast Distributions. Since sharpness is a measure of certainty, it is dependent on forecasting skill. The shape of a forecast distribution diagram also depends on the climatological frequency of the event being forecast. These two relationships have been modeled and are shown in Figure 4-2 (Boehm, 1976b). Skill in these examples is represented by the correlation of forecast probabilities with verifying observations and ranges from 0.2 (low skill) to 0.95 (high skill). These graphs are the same type as the graph in Figure 4-1, except the graph in Figure 4-1 was placed on its side and the order of probability values reversed. Notice the symmetry associated with distributions having a climatological probability of 0.5, and the skewness tendency as the climatological probability decreases; i.e., the skewness varies in proportion to the climatological frequency. Also, notice the high degree of sharpness corresponding with high skill, and near zero sharpness corresponding with low skill. Although these distributions are theoretical models and assume perfect reliability, they can be used as the ideal when subjectively evaluating forecast distribution diagrams for sharpness. Similar distributions for climatic probabilities greater than 0.5 would be a mirror image of those below 0.5.



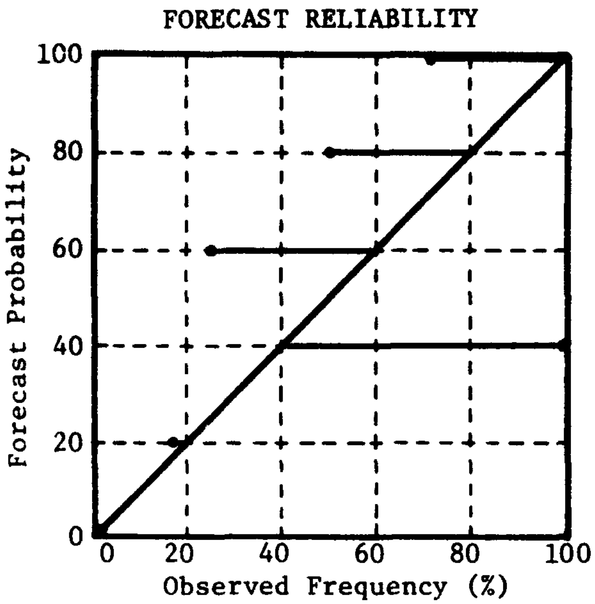
NOTE: 1. Graphs assume perfect reliability.

2. On individual graphs, abscissa is the forecast probability (0-100%), and ordinate is the relative frequency of forecasts.

3. Forecast correlation is the correlation between forecast and observed events.

Figure 4-2. Forecast distribution frequency graphs as a function of forecasting skill and the climatological frequency of the event.

FORECAST RELIABILITY



FORECAST VERIFICATION

OCCUR-RENCES	# OF FCSTS	OBSVD FREQ
5	7	71%
2	4	50%
1	4	25%
1	1	100%
1	6	17%
0	9	0%
Totals		Climo
10	31	32%

FORECAST DISTRIBUTION

Forecast Probability	1	2	3	4	5	6	7	8	9	10	11	12
100	1	0	1	1	1	1	0					
80	1	1	0	0								
60	1	0	0	0								
40	1											
20	0	0	1	0	0	0						
0	0	0	0	0	0	0	0	0	0			

Legend: 1-Represents event occurrences.
0-Represents nonoccurrences.

Figure 4-3. Example Reliability and Forecast Distribution Diagrams.

b. Reliability. Reliability is a measure of a forecaster's ability to accurately assign probability values. It reflects the degree that forecast probabilities resemble the observed frequency for each forecast probability interval. For example, an event would occur 80% of the time for a series of perfectly reliable 80% probability forecasts. Reliability does not measure skill, since always forecasting the climatological probability would be a perfectly reliable forecast (Sanders, 1963). However, it is a measure of how well forecasters know their skill limits. No single forecast can be judged as to its reliability; reliability can be evaluated only for a set of forecasts. "Perfect" reliability occurs when forecast probabilities are the same as observed frequencies for each probability interval throughout the range of probabilities (0-100%). "Zero" reliability occurs when all forecasts are exactly wrong; i.e., all forecasts were for values of either 0% or 100%, and the observed frequencies were the opposite. Thus, only 0% or 100% probability forecasts can be perfectly right or wrong. Intermediate values are only partially right or wrong.

(1) Reliability Diagrams. To measure reliability, graph the observed frequency for each forecast probability interval against the forecast value. Figure 4-3 is the reliability diagram that goes with the forecast set presented in Figure 4-1.

(2) If forecasts are perfectly reliable, plots of the observed frequency fall exactly on the diagonal line, commonly called the line of perfect reliability. Most plotted values of observed frequency in Figure 4-3 do not fall on this line. Horizontal lines were drawn from the diagonal to the plotted points to indicate their distance from the line of perfect reliability. We know most of these forecasts were not reliable, but now we must determine if these deviations were significant. A simple test for determining the significance of deviations from the line of perfect reliability is either to add or to subtract "one" from the number of events that occurred at the probability interval under investigation. Add if the plotted point is to the left of the diagonal; subtract if it is to the right. Recompute the observed frequency. If the line of perfect reliability falls between the actual and test values, the deviation is not considered significant. If the diagonal still is not reached, the deviation from perfect reliability is significant, and forecast performance needs improvement.

(a) To illustrate, consider forecast performance at the 100% interval. Adding "one" to the five occurrences raises the observed frequency to 86% (6/7) which is still less than 100%; thus, this deviation from perfect reliability is significant. Using this test for the remaining probability intervals shows the deviations at 80% and 60% are significant, and those at 40% and 20% insignificant.

(b) This test only tells us whether or not the deviation from perfect reliability is important, when there are a small number of occurrences involved. The test gives no information about how good or how bad the significant deviations are. This must be judged from the impact of unreliable forecasts on operational missions. Note, rare events will show large deviations, many of which will be classed as insignificant by using this test. Therefore, one should be cautious in applying this test when the climatic frequency is very low.

c. Over-underforecasting and Over-underconfidence. There are four special cases of deviations from perfect reliability: overforecasting, underforecasting, overconfidence, and underconfidence (Sanders, 1958). These are illustrated in Figure 4-4. In each case, deviations extend over all probability intervals and are identified by the hatched areas.

(1) Overforecasting results when a forecaster uses probability values that are too high compared to the observed frequency. All deviations on the same side of the line of perfect reliability indicates a problem exists, even if all of the deviations are not significant.

(2) Underforecasting occurs when the probability values used are too low compared to the observed frequency.

(3) Overconfidence results from trying to achieve greater sharpness than is warranted by forecasting skill. It is the excessive use of higher and lower probability values on the respective sides of the climatological frequency. This is very common with experienced forecasters in their early attempts at probability forecasting; it is considered to be a residual effect of categorical forecasting.

(4) Underconfidence results from understating the probability of occurrence of the event; i.e., hedging the forecast away from the extremes (0% and 100%) toward the climatological frequency. It is characteristic of individuals who are overly cautious and are not displaying their full forecasting abilities.

4-3. Controlling Sharpness and Reliability. The objective in probability forecasting is to achieve the optimum balance between sharpness and reliability. Excessive sharpness will show up as bias in reliability which can be corrected. However, an underestimate of skill (inadequate sharpness) can unknowingly exist and never be reflected in reliability measures (Hughes, 1965). Therefore, forecasters must not be content with perfect reliability. Remember that constant forecasts of the climatological probability will be perfectly reliable in the long run, but have zero sharpness and require no skill. Initial efforts in probability forecasting must concentrate on attaining acceptable reliability. Experiences of NWS indicate that forecasters can quickly adjust their biases, given timely feedback (Hughes, 1976a). Once the forecasts are consistently reliable, emphasis should shift toward maximizing sharpness, and then continually strive for the proper balance of the two.

a. Bias. The term "bias" is frequently used in conjunction with the four characteristics of over/underforecasting and over/underconfidence to indicate the magnitude and direction of the tendency to deviate from perfect reliability. A value of bias can be determined for each forecast probability interval, as well as for the entire set of forecasts overall. The former is called interval bias; the latter, overall bias.

(1) Interval bias. Bias for each probability interval is computed by subtracting the observed frequency from the probability value of the corresponding forecast interval (Hughes, 1976a). For example, biases for the example given in Figure 4-3 are: 100-71% = +29%, 80-50% = +30%, 60-25% = +35%, = 40-100% = -60%, 20-17% = 3%, and 0-0% = 0%. The sign of the bias value indicates the type of bias, i.e., positive values reflect overforecasting; negative values, underforecasting. The magnitude of interval bias indicates the percentage difference between the

observed frequency and perfect reliability or, the reliability error. The significance of interval bias depends on the number of forecasts in each interval. A large bias in only one interval containing a small number of forecasts is not significant, unless adjacent intervals have the same kind of bias. Further, small biases that alternate in type (sign) with increasing or decreasing probability are usually the result of sampling error. However, a series of biases of the same type, even for small values, indicates undesirable trends.

(2) Overall bias. One method to make a quick check for reliability errors in a set of forecasts is to calculate the overall bias (B) by using the equation.

$$B = \frac{\Sigma P - O}{N} \tag{4-1}$$

where 0 is the total number of event occurrences in the set of forecasts, N is the total number of forecasts made, and P is the sum of all the probability values used in the set. The latter can be computed by adding all individual probability values, or by multiplying the probability times the number of forecasts in each interval and then adding (remember to use decimal values of probabilities in all formulas). The latter method is recommended because it is easier and quicker; Table 4-1 demonstrates this computational method. Another equation for overall bias is (Hughes, 1976b):

$$B = \frac{\Sigma P - O}{O} \tag{4-2}$$

While both equations are proper, 4-1 is used here to be compatible with the method used for interval bias and to place finite limits on the range of B encountered.

(a) The four examples shown depict the relationship between interval bias and overall bias and demonstrate how bias can be used to determine reliability. For example, the set of forecasts with overforecasting have a positive bias in all but one interval, and an overall positive bias of .1 or 10%. Since this is a pure case of overforecasting where all interval biases are plus 10%, the obvious solution for achieving perfect reliability would be to move the probability values of all the forecasts down one interval. In other words, the forecaster should be instructed to reduce forecast probabilities by 10% in every interval for his next set of forecasts. Underforecasting is exactly the opposite problem. Here the forecaster should be told to raise his probability values by 10% in future forecasts. Overconfidence is a combination of over and underforecasting. In this example, the forecasts were 10% too high above sample climatology (50 events/100 forecasts = 50%) and 10% too low below sample climatology. To improve, the forecaster should reduce his forecast probabilities above the climatological probability by 10% and increase those below climatology by 10%. Underconfidence is the opposite of overconfidence and, when diagnosed, should be corrected by making the opposite corrections as for the overconfidence example. Refer back to Figure 4-4 to see these reliability biases in graphical format.

(b) Absence of overall bias does not necessarily mean the absence of reliability problems (Hughes, 1976a). In Table 4-1, notice that overall bias for both over and underconfidence is zero. This is because overall bias is actually the weighted average of positive and negative interval biases, which, in this example, cancels values of equal but opposite sign. Therefore, a forecaster should inspect interval bias as well as overall bias, because large interval biases could exist even though overall bias is zero. On the other hand, an overall bias indicates a reliability problem, and the type of bias (overforecasting or underforecasting).

b. Figure 4-5 shows additional examples of the use of sharpness and reliability diagrams to evaluate probability forecasts.

(1) In the first example (overforecasting), a positive bias of 20% occurred in the 100% probability interval. By using the significance test from para 4-2b(2), we see this is on the borderline for classification as significant; i.e., the test value equals perfect reliability. However, since this is the only interval with a deviation from perfect reliability, one should seek to correct it. A possible explanation is that the deviation occurred because either forecast skill or the state-of-the-forecast-art was exceeded. The forecasts were for 100% probability, while the observed frequency was only 80%. If an 80% probability had been assigned to these five forecasts, they would have been perfectly reliable. Consequently, the forecaster should be instructed to avoid using 100% probabilities in future forecasts unless he is certain. This forecaster should also be instructed to improve sharpness, i.e., to try to better identify those cases when high and low probabilities are justified.

(2) In the underforecasting example, significance tests show that the deviation for the 80% probability interval is significant, and deviations at the other intervals are borderline. Even if all deviations were classified as insignificant, the forecaster should be concerned, because all the biases have the same sign. To improve reliability, this forecaster should use a probability value one interval higher in future forecasts. Too many probabilities are being assigned in the middle intervals. In summary, this problem is the inability to recognize those cases when the threshold is met (indicated by a "1" for verification purposes).

(3) The overconfidence example indicates that forecasting skill was exceeded. Note that this forecaster has a good sharpness pattern—25 of his 31 forecasts were for 0% or 100%. The underconfidence example indicates an understatement of forecast skill; most forecast probabilities are grouped around climatology (54%), i.e., sharpness is bad. The examples given in Table 4-1 and Figure 4-5 were designed to show the mechanics of using bias to improve reliability. In actual practice, solutions will not be as clear. Sharpness and reliability problems will be mixed, and sampling problems (noise) can be quite large in small data samples.

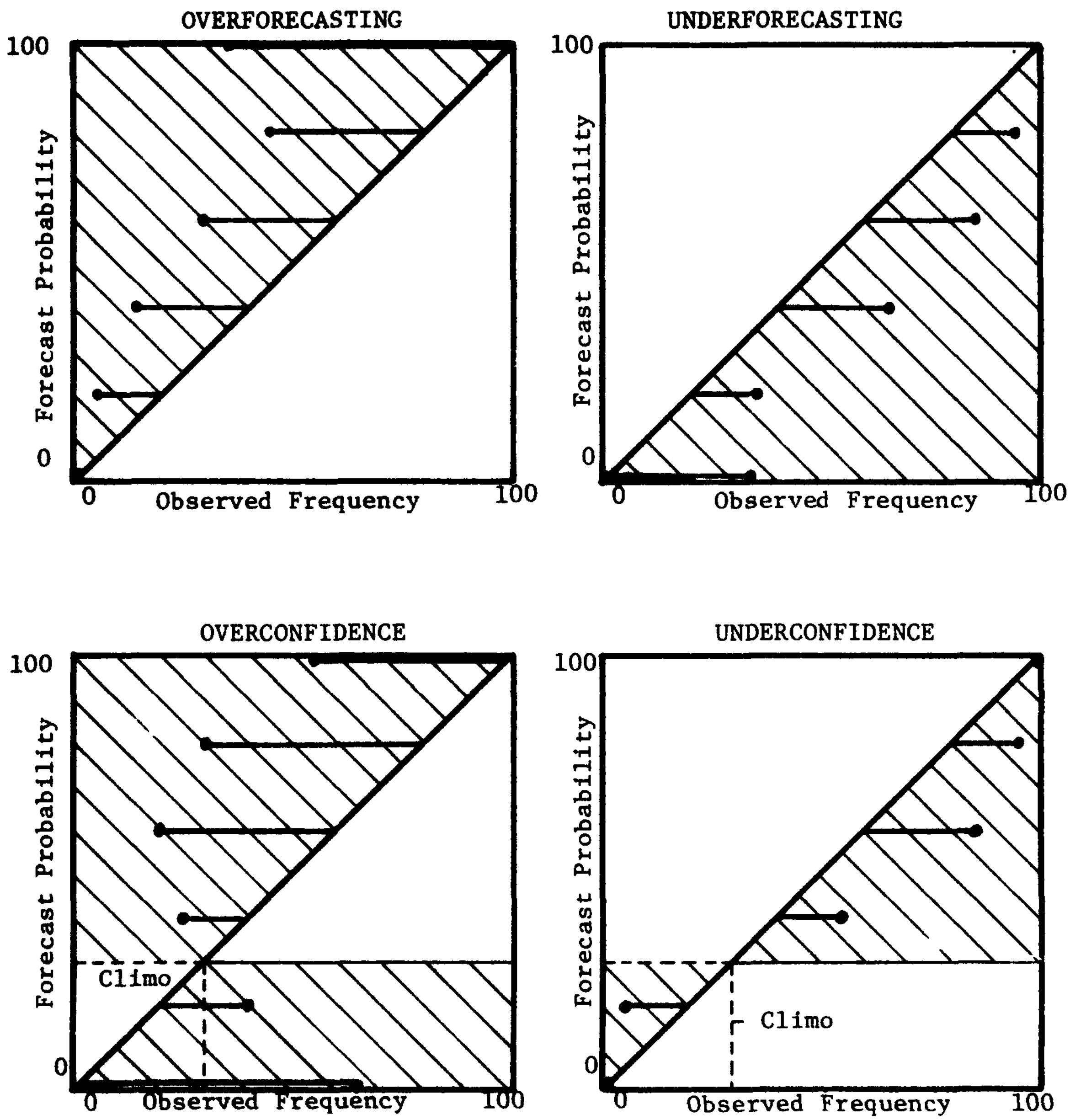


Figure 4-4. Over and Underforecasting and Over and Underconfidence.

Table 4-1. Using Bias Measures to Improve Reliability.

Definitions

P = Probability value for each interval

N = Number of forecasts for each interval

O = Number of event occurrences

O/N = Observed frequency

P-O/N = Interval bias

OVERFORECASTING					
P	N	PxN	O	O/N	P-O/N
1.0	20	20	18	.9	+.1
.9	10	9	8	.8	+.1
.8	10	8	7	.7	+.1
.7	10	7	6	.6	+.1
.6	10	6	5	.5	+.1
.4	10	4	3	.3	+.1
.3	10	3	2	.2	+.1
.2	10	2	1	.1	+.1
.1	10	1	0	0	+.1
.0	0	0	-	-	-
All	100	60	50		
Overall Bias $B = \frac{60-50}{100} = +.10$					

UNDERFORECASTING					
P	N	PxN	O	O/N	P-O/N
1.0	0	0	-	-	-
.9	10	9	10	1.0	-.1
.8	10	8	9	.9	-.1
.7	10	7	8	.8	-.1
.6	10	6	7	.7	-.1
.4	10	4	5	.5	-.1
.3	10	3	4	.4	-.1
.2	10	2	3	.3	-.1
.1	10	1	2	.2	-.1
.0	20	0	2	.1	-.1
All	100	40	50		
Overall Bias $B = \frac{40-50}{100} = -.10$					

NOTE: Probability interval of .5 was omitted for simplicity.

OVERCONFIDENCE					
P	N	PxN	O	O/N	P-O/N
1.0	20	20	18	.9	+.1
.9	15	13.5	12	.8	+.1
.8	10	8	7	.7	+.1
.7	5	3.5	3	.6	+.1
.6	0	0	0	-	-
.4	0	0	0	-	-
.3	5	1.5	2	.4	-.1
.2	10	2	3	.3	-.1
.1	15	1.5	3	.2	-.1
.0	20	0	2	.1	-.1
All	100	50	50		
Overall Bias $B = \frac{50-50}{100} = .00$					

UNDERCONFIDENCE					
P	N	PxN	O	O/N	P-O/N
1.0	0	0	-	-	-
.9	5	4.5	5	1.00	-.10
.8	10	8.0	9	.90	-.10
.7	20	14.0	16	.80	-.10
.6	15	9.0	10	.67	-.07
.4	15	6.0	5	.33	+.07
.3	20	6.0	4	.20	+.10
.2	10	2.0	1	.10	+.10
.1	5	.5	0	.00	+.10
.0	0	0	-	-	-
All	100	50	50		
Overall Bias $B = \frac{50-50}{100} = .00$					

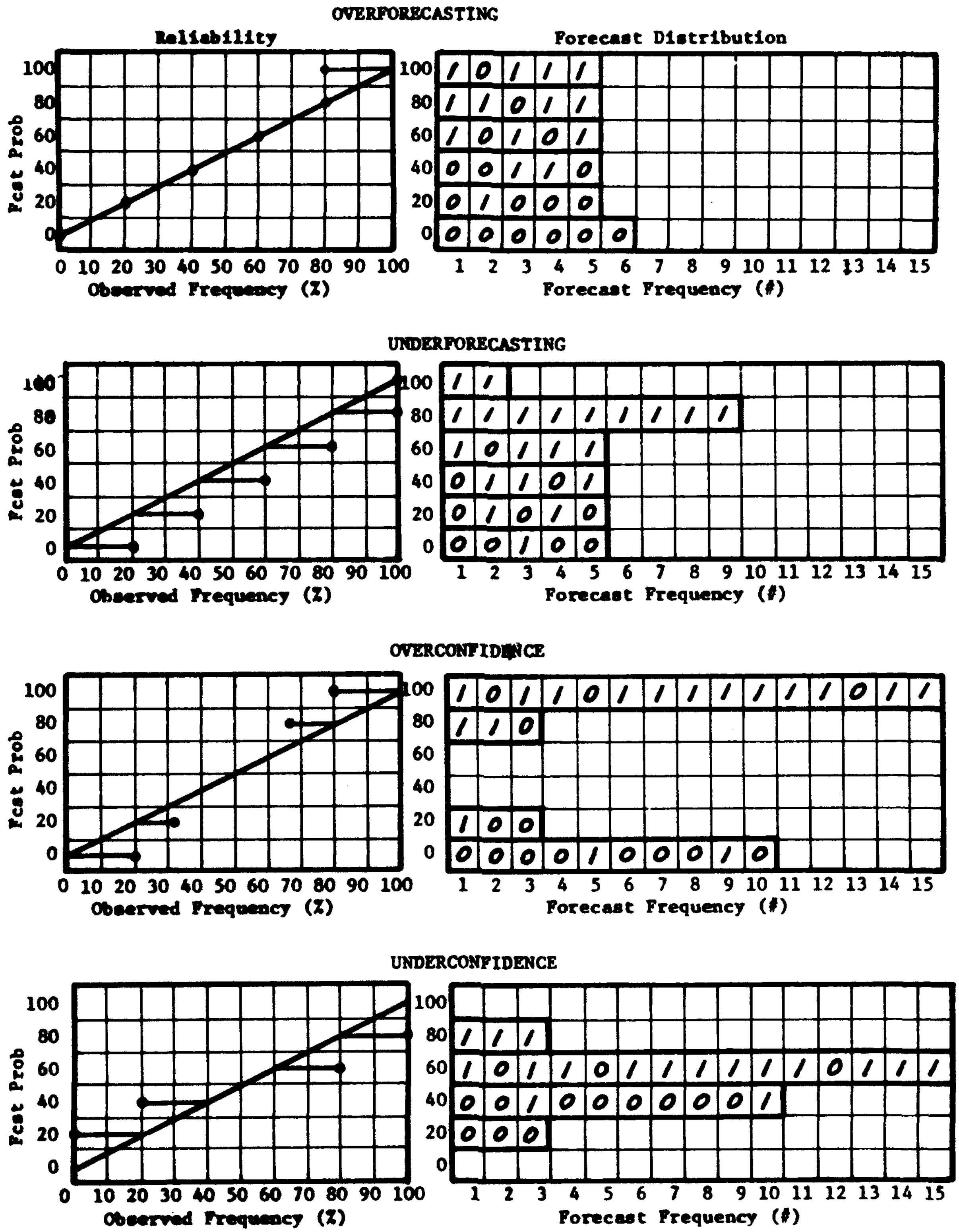


Figure 4-5. Examples of Over and Underforecasting and Over and Underconfidence.

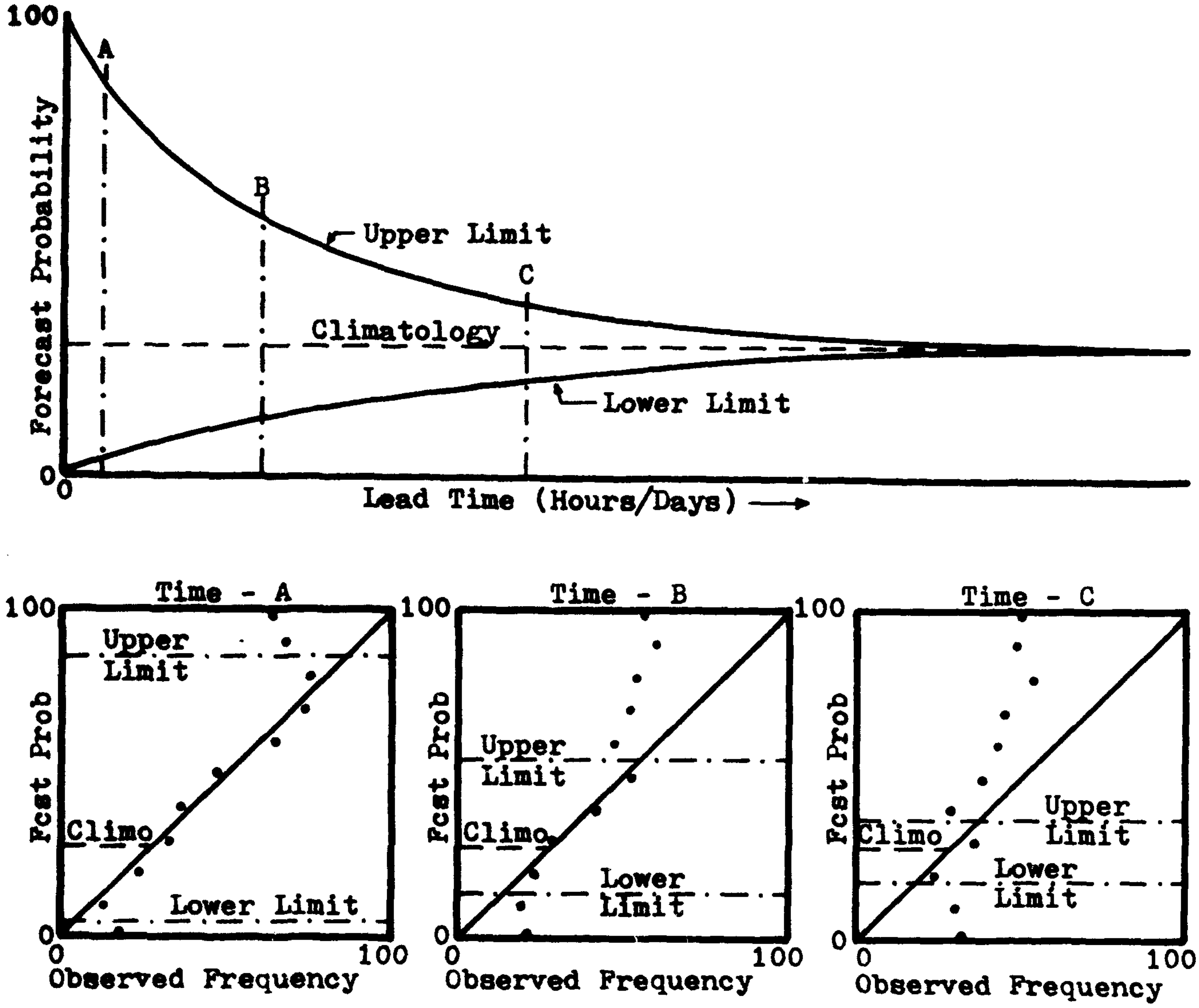


Figure 4-6. Limits of Reliable Forecasts Versus Lead Time.

c. Establishing and Using Reliability Standards. By using the principle that forecasting skill decreases with increasing length of time (or lead time) of forecasts, Hughes (1965, 1966, 1967b) has shown that this reduction of skill also shrinks the useable range of reliable probabilities available to the forecaster. The illustration at the top of Figure 4-6 depicts this concept. For forecasts with short lead time, it is usually possible to use the full range of probability values 0 through 100% and still achieve good reliability. However, as lead time increases (and skill or state-of-the-art decreases) the upper and lower limits of reliable forecasts shrink and converge to the climatological probability. The exact shape of the curves and the point at which they converge to climatology will vary with the event, its climatic frequency, the forecasting state-of-the-art for the event, and with the individual's skill.

(1) The three reliability diagrams at the bottom of Figure 4-6 depict how the top diagram might be derived (Hughes, 1965). First, standard reliability diagrams are plotted for forecasts of the given event; separate diagrams are plotted for selected lead times. The next step is to identify the upper and lower limits of acceptable reliability. By using a standard agreeable to the customer, determine the upper and lower forecast probability values which separate the reliable and unreliable areas on the diagram. In this example a bias of greater than ± 5% deviation from perfect reliability was used to flag the unreliable areas. Horizontal lines depicting upper and lower limits of reliability were drawn at the forecast probability value above or below which deviations exceeded the standard. The probability values which separate the areas were plotted on the top diagram at the appropriate lead time. Smoothed upper and lower limit curves were then drawn connecting the plotted points. In actual practice, most units will not issue forecasts which have a lead time extending out to the time where the upper and lower limits converge. However, if such a diagram is required, the right hand portion of the diagram may have to be extrapolated. This type of diagram has the advantage of showing the cut-off point beyond which no skill exists, and when climatology should be used as the forecast.

(2) Consider a short range forecast for ceilings below 5,000 feet with a lead time of one to three hours. There would be many times that the forecaster would be certain that the event would or would not occur; consequently, probabilities of 0 to 100% could be used reliably. Further, there are times when these values could be used with much longer lead times. But, for forecasts out to 24 to 36 hours, skill and reliability limits would most likely be exceeded if probabilities of 0 and 100% were used frequently.

(3) Next, consider a forecast for a rare event such as tornadoes. There should be very few times that forecasters would use 100% probability, and those times most likely would occur only after a tornado has been sighted or detected on radar. Use of high probabilities would decrease very rapidly with lead time and converge to climatology (which is also extremely low in this case) for a lead time of a few hours. Thus, one seldom uses high probabilities to forecast rare events. When values larger than zero are used, they should not be substantially greater than the climatological probability except for very short lead times.

(4) The initial bias of almost every forecaster inexperienced in probability forecasting is one of overestimating the degree of skill possessed. Not fully realizing their limits, forecasters generally use high and low probabilities values too frequently especially for the longer lead times, resulting in poor reliability. (Hughes, 1976d). Reliability diagrams with upper and lower limits added to them can greatly aid in minimizing bias problems by controlling the use of unreliable probability values. In operational use, supervisors can instruct forecasters not to use values outside acceptable limits unless fully justified by a well organized and easily predicted synoptic situation.

(5) The same information contained in reliability diagrams can be derived by inspection of the biases for each probability interval. Reliability limits can then be obtained for each forecast event and each time period. Limits derived from overall unit performance are useful for briefing customers. Reliability limits should be determined for new forecasters to enable them to rapidly overcome their biases. The larger the data base, the more reliable the information will be. Individual reliability limits should be reexamined periodically since forecasting skill should increase.

(6) Reliability limits will be required for each forecast event. The standard used to determine upper and lower limits of acceptable reliability should be dictated by the effect of unreliable forecasts on the operation in which they will be used. However, in the absence of reliability requirements from the customer, a recommended standard is that the bias be within ± 5% of the forecast probability value. A unit's standard should apply to intermediate probability intervals as well as the upper and lower limits. Finally, it may be necessary for a unit to determine reliability limits for each season.

(7) Similar procedures were used by one region of NWS to establish a policy for their precipitation probability forecasts. Forecasters were instructed not to use probabilities beyond the limits listed in Table 4-2, unless unusually favorable and well defined conditions would justify their use. This guidance was provided during their early experience in precipitation probability forecasting and is still considered reasonable for this event. These figures are based on average precipitation climatological frequencies. In drier parts of the country, both limits would be reduced somewhat; in wetter areas, they would be increased (Hughes, 1976a).

Table 4-2. Limits for Reliable Precipitation Probabilities (NWS).

Valid Period (Hrs)	Probability Limits (%)
0-12	0-100
12-24	2-80
24-36	5-70
36-48	10-50

d. Evaluation Feedback. Timely feedback of verification results is extremely important in probability forecasting. Forecasters must know what their problem areas are. This is especially true for inexperienced forecasters just learning the procedures, and for experienced forecasters producing forecasts for a new event or a new station. In these cases, reliability is initially erratic. Forecasters can generally achieve acceptable reliability, if they are given timely verification feedback (Hughes, 1965). As a rough rule-of-thumb, reasonably good reliability can be expected by the time a forecaster has made 50 to 100 forecasts that involve occurrences of an event. Once the ability to maintain acceptable reliability has been achieved, efforts should concentrate on improving sharpness. Periodic feedback will still be required to insure the proper balance between sharpness and reliability.

(1) The minimum data for evaluating probability forecasts are: a table listing the probability intervals used to verify the forecasts; the corresponding number of forecasts, event occurrences, observed frequency and bias for each interval; appropriate totals, and overall bias. Examples of these data were given in Table 4-1. Verification results will be needed for each forecast event, each category if the forecast is for more than two categories (there are always at least two; e.g., rain or no rain or ceiling >or $\leq$1,000 ft), and for a representative number of forecasts. This information should be prepared for each forecaster and for the unit overall. Monthly verification should be maintained to identify trends. However, it may be necessary to combine data (number of forecasts and number of event occurrences by probability interval) for several months in order to have enough cases for meaningful evaluations.

(2) As an example, consider an evaluation of the probability forecasts shown in Figure 4-3. This diagram illustrates the reliability and sharpness one might obtain from an evaluation of a set of 31 forecasts issued once daily for the occurrence of flying weather below 3000 ft and/or 3 miles.

(a) This set of forecasts exhibits a fairly good degree of sharpness, i.e., 16 of 31 forecasts were in either the 0 or 100% probability intervals with another 10 in adjacent intervals (4 in 80% and 6 in 20%). Note that only one forecast was in the interval closest to sample climatology (32%), i.e., zero sharpness was not a problem. If these forecasts had exhibited perfect sharpness, all would have fallen in either the 0 or 100% intervals. Additionally, if the forecasts were all perfectly accurate, the forecast probabilities would have been distributed in those two intervals in proportion to the number of observed and not observed cases; i.e., the 10 event occurrences would all be in the 100% interval and all 21 nonoccurrences in the 0% interval.

(b) The reliability deviations at 100%, 80%, and 60% are significant. All forecast probabilities of 60% and greater were considerably larger than the observed frequencies. In order to improve his reliability, the forecaster should reduce all of his probability estimates that are above 60% by 10% for his next series of forecasts.

(3) General performance and specific problems can be more easily identified during initial phases by studying forecast distribution and reliability diagrams. All the data required to plot these diagrams are contained in the recommended table. Once the forecasters achieve proficiency in analyzing the data, diagrams for individual forecasters could be eliminated.

4-4. Brier Probability Score (PS). The Brier Probability Score is used to quantify the overall quality of probability forecasts. Its advantages and disadvantages are listed below, followed by the paragraph in this pamphlet which addresses each one. The advantages of using the Brier Score to evaluate a set of probability forecast are: one number is given which includes sharpness and reliability) (paragraph 4-4a); the score cannot be "played" (paragraph 4-4c); and the score can be used to compare different forecast systems (paragraph 4-4f).The disadvantages of the Brier Score are: it does not indicate if a set of forecasts are bad due to sharpness of reliability error (paragraph 4-4b); it is affected by the event's climatology (paragraph 4-4d); it is affected by the number of event categories (paragraph 4-4d); and a score for "zero skill" cannot be computed (paragraph 4-4e).

a. Computation. The equation for computing the Brier Probability Score is (Panofsky and Brier, 1965):

$$PS = \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{N} (R_{ij} - D_{ij})^2 \tag{4-3}$$

where K is the number of categories (2 or more)
N is the number of forecasts being evaluated
R_{ij} is the probability given for the i th forecast for the occurrence of category j weather
D_{ij} equals one if category j occurred for the i th forecast, otherwise $D_{ij} = 0$
PS is the Brier Score. A perfect score is 0.0. The worst possible score is 2.0.

For those unfamiliar with the mathematical symbology, Attachment 4 provides a complete explanation.This general equation may be used to compute Brier Score for forecasts of a number of categories ($K \geq 2$). For verification purposes, an "observed" probability of either 1.0 (event occurred) or 0.0 (event did not occur) is assigned to D (Sanders, 1958). Thus, the Brier Score is the average of the square of the differences between the forecast and "observed" probabilities. Since the score ranges from 0 (perfect) to 2 (worst possible), another aid to understanding its meaning is to think of the score in terms of penalty points; i.e., the worse the forecast, the larger the penalty (Hughes, 1965).

(1) If one is concerned only with two categories (K = 2), the general equation can be greatly simplified. For a two category forecast, the event either occurs or it doesn't; e.g., rain or no rain. The probability that the event will not occur equals one minus the probability that the event will occur. In the terminology used in the general Brier Score equation, $R_{i2} = 1 - R_{i1}$ and $D_{i2} = 1 - D_{i1}$. Substituting in the general equation, we obtain the Brier Score equation for forecasts containing only two categories:

$$PS = \frac{2}{N} \sum_{i=1}^{N} (R_i - D_i)^2 \tag{4-4}$$

where definitions are the same as in the general equation. This equation shows that the contribution to the Brier Score from one category is exactly equal to the contribution of the other. Therefore, the Brier Score for a two category forecast may be obtained by evaluating only a single category. It doesn't matter which one. For example, consider a forecast for 90% probability of rain. By using equation 4-4, the Brier Score for that one forecast would be calculated as follows:

If it rained, $R_i = .9$, $N = 1$, and $D_i = 1$; therefore,

$$PS = 2\,(.9-1)^2 = .02$$

If no rain occurred, D = 0; therefore,

$$PS = 2\,(.9-0)^2 = 1.62$$

In the first case, the forecast was nearly completely right: 90% probability of rain and it occurred. The penalty for the near miss was only 0.02. But in the second case, the error was large (nearly completely wrong). Here the forecast probability was 90%, whereas the observed probability was 0%. Consequently, the penalty is very high - near the maximum of 2.

(2) Rather than expanding the equation in traditional mathematical form and substituting values for the variables, a table can be used to perform the computations very quickly and simply.

(a) Table 4-3 illustrates how the forecasts may be recorded and the Brier Score computed for a two category forecast by using equation 4-4 directly. In actual practice, the columns labeled "Fcst%" and "Verification" could be omitted, since they only show how the values in columns labeled "R_i" and "D_i", respectively, were derived. The last column contains the penalties associated with each forecast. They are added, multiplied by 2 (since this is only one of two categories), and averaged by dividing by the number of forecasts (20) to obtain the Brier Score for the entire set. The overall bias shows underforecasting. Interval bias cannot be computed, unless the forecasts are grouped by interval.

TABLE 4-3. SAMPLE FORMAT FOR RECORDING AND COMPUTING TWO CATEGORY BRIER SCORE.

i	Valid Date (12Z)	Fcst %	R_i	Verification	D_i	$(R_i - D_i)$	$(R_i - D_i)^2$
1	10	0	.0	NO	0.	0.0	0.00
2	11	80	.8	YES	1.	−0.2	0.04
3	12	0	.0	YES	1.	−1.0	1.00
4	13	0	.0	NO	0.	0.0	0.00
5	14	0	.0	NO	0.	0.0	0.00
6	15	0	.0	NO	0.	0.0	0.00
7	16	80	.8	YES	1	−0.2	0.04
8	17	0	.0	NO	0.	0.0	0.00
9	18	0	.0	NO	0.	0.0	0.00
10	19	0	.0	NO	0.	0.0	0.00
11	20	0	.0	NO	0.	0.0	0.00
12	21	0	.0	NO	0.	0.0	0.00
13	22	60	.6	NO	0.	0.6	0.36
14	23	20	.2	NO	0.	0.2	0.04
15	24	30	.3	NO	0.	0.3	0.09
16	25	60	.6	NO	0.	0.6	0.36
17	26	80	.8	YES	1.	−0.2	0.04
18	27	90	.9	YES	1.	−0.1	0.01
19	28	70	.7	YES	1.	−0.3	0.09
20	29	20	.2	YES	1.	−0.8	0.64
TOTALS			5.9		7		2.71

$$PS = \frac{2}{N} \sum_{i=1}^{N} (R_i - D_i)^2 = \frac{2}{20} \times 2.71 = 0.27 \qquad \text{OVERALL BIAS, } B = \frac{5.9 - 7}{20} = -0.055$$

(b) The Brier Score for a four category forecast is shown in Table 4-4. The unnecessary columns used in Table 4-3 were eliminated to show the minimum information required to compute the score. The last two columns in each category are shown only to indicate how a running account of the Brier Score may be accomplished. "Penalty Sum" is the running total or accumulated penalties and "PS $_{1, 2, 3, 4}$" is the partial Brier Score for all forecasts (i). The total Brier Score for all categories in the set is simply the sum of the scores for each category.

(3) If daily computations of the Brier Score are not needed, the procedures can be shortened even more. Although not as precise as using the equation directly, forecast probabilities can be grouped into fixed intervals as demonstrated earlier in the discussion of sharpness and reliability (Table 4-1). In this case, the differences between the forecast and "observed" probabilities (0 or 1) would be a set of constants. This feature allows one to precompute and square all the possible differences between the two probabilities and prepare a table of partial Brier Scores (or penalty points). Such a table is given in Attachment 5. The word "partial" is used because penalties for all occurrences and nonoccurrences must be added and then divided by the number of forecasts involved to obtain the Brier Score for the category being evaluated. If the forecast is for two categories, multiply by two; otherwise, the Brier Scores for all categories must be summed to obtain a total score.

(a) Table 4-5 illustrates how the data from the first two columns of each category in Table 4-4 may be grouped into probability intervals. Brier Scores were computed by using data from Table A5-2 (Atch 5). In each category, penalties for occurrences were extracted first; then those for nonoccurrences were derived. The values were added and the sum divided by the number of forecasts (20) to obtain the Brier Score for each category. A total Brier Score was found by summing values for the four categories.

(b) Interval bias was added to make the summary all inclusive. This summary includes all the information needed to plot reliability and forecast distribution diagrams for each category. Figure 4-7 shows the corresponding diagrams. The data used in this series of tables and diagrams were chosen to represent results that might occur in evaluating ceiling forecasts. Note how sparsity of data in some probability intervals makes the evaluation difficult.

(4) Admittedly, the Brier Score could be based on something other than sharpness and reliability. We selected this partitioning because it provides us the information we want.

(5) If a unit wants to automate Brier Score computations, contact AWS/DNT for assistance.

b. Relationship of Brier Score to Sharpness and Reliability. Figure 4-8 illustrates how the Brier Score varies with forecast probability and observed frequency. For any reasonable and likely reliability, the range of the score is approximately 0 to 0.6 rather than 0 to 2.0. The system encourages reliability, since the lowest score for any observed frequency is at the equivalent forecast probability (i.e., perfect reliability). Forecasts of 50% probability yield a poor score (.5) regardless of the reliability, while the greatest penalties for poor reliability are with very high and very low forecast probabilities (Hughes, 1965). Although the lowest scores are at zero observed frequency for forecast probabilities below 50%, and at 100% observed frequency for higher probabilities, sharpness is encouraged because the best overall scores are found at the extremes (Hughes, 1967a). Thus, the Brier Score provides a combined measure of reliability and the ability to move forecasts away from 50% probability (sharpness) (Hughes, 1965). The fact that the focal point for measuring sharpness is 50% probability, instead of climatology, is a deficiency which must be considered when interpreting the score. Examples showing penalty points and overall Brier Scores for various combinations of reliability and sharpness are illustrated in Table 4-6.

(1) The first example shows a set of forecasts with perfect reliability, but a constant number of forecasts in each probability interval (poor sharpness). Note how the penalties for occurrences and nonoccurrences are reciprocals, and that the maximum total penalty occurs at the center of the probability intervals (50% probability was omitted intentionally to simplify the next two examples). Since reliability is perfect, the resultant Brier Score is due solely to poor sharpness.

Table 4-4. Example Verification for a Four Category Forecast.

Fcst # (i)	Obsvd Cat	CATEGORY 1 Fcst Prob (R_{11})	Obsvd Prob (D_{11})	Penalty $(R_{11}-D_{11})^2$	Penalty Sum	PS_1	CATEGORY 2 Fcst Prob (R_{12})	Obsvd Prob (D_{12})	Penalty $(R_{12}-D_{12})^2$	Penalty Sum	PS_2	CATEGORY 3 Fcst Prob (R_{13})	Obsvd Prob (D_{13})	Penalty $(R_{13}-D_{13})^2$	Penalty Sum	PS_3	CATEGORY 4 Fcst Prob (R_{14})	Obsvd Prob (D_{14})	Penalty $(R_{14}-D_{14})^2$	Penalty Sum	PS_4
1	4	.0	0	.00	.00	.000	.0	0	.00	.00	.000	.1	0	.01	.01	.010	.9	1	.01	.01	.010
2	4	.0	0	.00	.00	.000	.0	0	.00	.00	.000	.0	0	.00	.01	.005	1.0	1	.00	.01	.005
3	4	.0	0	.00	.00	.000	.0	0	.00	.00	.000	.0	0	.00	.01	.003	1.0	1	.00	.01	.003
4	4	.0	0	.00	.00	.000	.1	0	.01	.01	.003	.2	0	.04	.05	.013	.7	1	.09	.10	.025
5	3	.0	0	.00	.00	.000	.1	0	.01	.02	.004	.8	1	.04	.09	.018	.1	0	.01	.11	.022
6	2	.1	0	.01	.01	.002	.8	1	.04	.06	.010	.1	0	.01	.10	.017	.0	0	.00	.11	.018
7	3	.0	0	.00	.01	.001	.8	0	.64	.70	.100	.2	1	.64	.74	.106	.0	0	.00	.11	.016
8	4	.0	0	.00	.01	.001	.0	0	.00	.70	.088	.1	0	.01	.75	.094	.9	1	.01	.12	.015
9	4	.0	0	.00	.01	.001	.0	0	.00	.70	.078	.0	0	.00	.75	.083	1.0	1	.00	.12	.013
10	3	.0	0	.00	.01	.001	.0	0	.00	.70	.070	1.0	1	.00	.75	.075	.0	0	.00	.12	.012
11	4	.0	0	.00	.01	.001	.0	0	.00	.70	.064	.0	0	.00	.75	.068	1.0	1	.00	.12	.011
12	4	.0	0	.00	.01	.001	.0	0	.00	.70	.058	.0	0	.00	.75	.063	1.0	1	.00	.12	.010
13	2	.1	0	.01	.02	.002	.8	1	.04	.74	.057	.1	0	.01	.76	.058	.0	0	.00	.12	.009
14	1	.4	1	.36	.38	.027	.5	0	.25	.99	.071	.1	0	.01	.77	.055	.0	0	.00	.12	.009
15	3	.0	0	.00	.38	.025	.1	0	.01	1.00	.067	.8	1	.04	.81	.054	.1	0	.01	.13	.009
16	4	.0	0	.00	.38	.024	.0	0	.00	1.00	.063	.8	0	.64	1.45	.091	.2	1	.64	.77	.048
17	4	.0	0	.00	.38	.022	.0	0	.00	1.00	.059	.0	0	.00	1.45	.085	1.0	1	.00	.77	.045
18	4	.0	0	.00	.38	.021	.1	0	.01	1.01	.056	.7	0	.49	1.94	.108	.2	1	.64	1.41	.078
19	3	.1	0	.01	.39	.021	.5	0	.25	1.26	.066	.2	1	.64	2.58	.136	.2	0	.04	1.45	.076
20	4	.0	0	.00	.39	.020	.1	0	.01	1.27	.064	.2	0	.04	2.62	.131	.7	1	.09	1.54	.077
TOTALS		.7	1	.39			3.9	2	1.27			5.4	5	2.62			10.0	12	1.54		
B_j		$B_1 = (.7-1)/20 = -.015$					$B_2 = (3.9-2)/20 = +.095$					$B_3 = (5.4-5)/20 = +.02$					$B_4 = (10-12)/20 = -.10$				
PS_j		$PS_1 = .39/20 = .020$					$PS_2 = 1.27/20 = .064$					$PS_3 = 2.62/20 = .131$					$PS_4 = 1.54/20 = .077$				
PS		Total Brier Score for All Categories, $PS = PS_1 + PS_2 + PS_3 + PS_4 = .020 + .064 + .131 + .077 = .292$																			

Table 4-5. Example Verification Summary for a Four Category Forecast.

	FCST PROB (R_{ij})	# FCST (n)	OCCURRENCES (D_{ij} = 1) #	OCCURRENCES (D_{ij} = 1) PENALTY	OBSVD FREQ	INTERVAL BIAS	PROB SUM $n(R_{ij})$	NON OCCURRENCES (D_{ij} = 0) #	NON OCCURRENCES (D_{ij} = 0) PENALTY	PENALTY SUM
CATEGORY 1	1.0									
	.9									
	.8									
	.7									
	.6									
	.5									
	.4	1	1	.36	1.00	-.60	0.4	0	.0	.36
	.3									
	.2									
	.1	3	0	.00	.00	+.10	0.3	3	.03	.03
	.0	16	0	.00	.00	.00	0.0	16	.00	.00
TOTAL/AVE		20	1	.36	.05		0.7	19	.03	.39
BIAS (B_1) PS$_1$		B_1 = (.7-1)/20 = -.015 PS$_1$ = .39/20 = .02								
CATEGORY 2	1.0									
	.9									
	.8	3	2	.08	.67	+.13	2.4	1	.64	.72
	.7									
	.6									
	.5	2	0	.00	.00	+.50	1.0	2	.50	.50
	.4									
	.3									
	.2									
	.1	5	0	.00	.00	+.10	.5	5	.05	.05
	.0	10	0	.00	.00	.00	.0	10	.00	.00
TOTAL/AVE		20	2	.08	.10		3.9	18	1.19	1.27
BIAS (B_2) PS$_2$		B_2 = (3.9-2)/20 = +.095 PS$_2$ = 1.27/20 = .064								
CATEGORY 3	1.0	1	1	.00	1.00	.00	1.0	0	.00	.00
	.9									
	.8	3	2	.08	.67	+.13	2.4	1	.64	.72
	.7	1	0	.00	.00	+.70	.7	1	.49	.49
	.6									
	.5									
	.4									
	.3									
	.2	4	2	1.28	.50	-.30	.8	2	.08	1.36
	.1	5	0	.00	.00	+.10	.5	5	.05	.05
	.0	6	0	.00	.00	.00	.0	6	.00	.00
TOTAL/AVE		20	5	1.36	.25		5.4	15	1.26	2.62
BIAS (B_3) PS$_3$		B_3 = (5.4-5)/20 = +.02 PS$_3$ = 2.62/20 = .131								
CATEGORY 4	1.0	6	6	.00	1.00	.00	6.0	0	.00	.00
	.9	2	2	.02	1.00	-.10	1.8	0	.00	.02
	.8									
	.7	2	2	.18	1.00	-.30	1.4	0	.00	.18
	.6									
	.5									
	.4									
	.3									
	.2	3	2	1.28	.67	-.47	.6	1	.04	1.32
	.1	2	0	.00	.00	+.10	.2	2	.02	.02
	.0	5	0	.00	.00	.00	.0	5	.00	.00
TOTAL/AVE		20	12	1.48	.60		10.0	8	.06	1.54
BIAS (B_4) PS$_4$		B_4 = (10-12)/20 = -.10 PS$_4$ = 1.54/20 = .077								
PS$_{All}$		PS = PS$_1$ + PS$_2$ + PS$_3$ + PS$_4$ = .02 + .064 + .131 + .077 = .292								

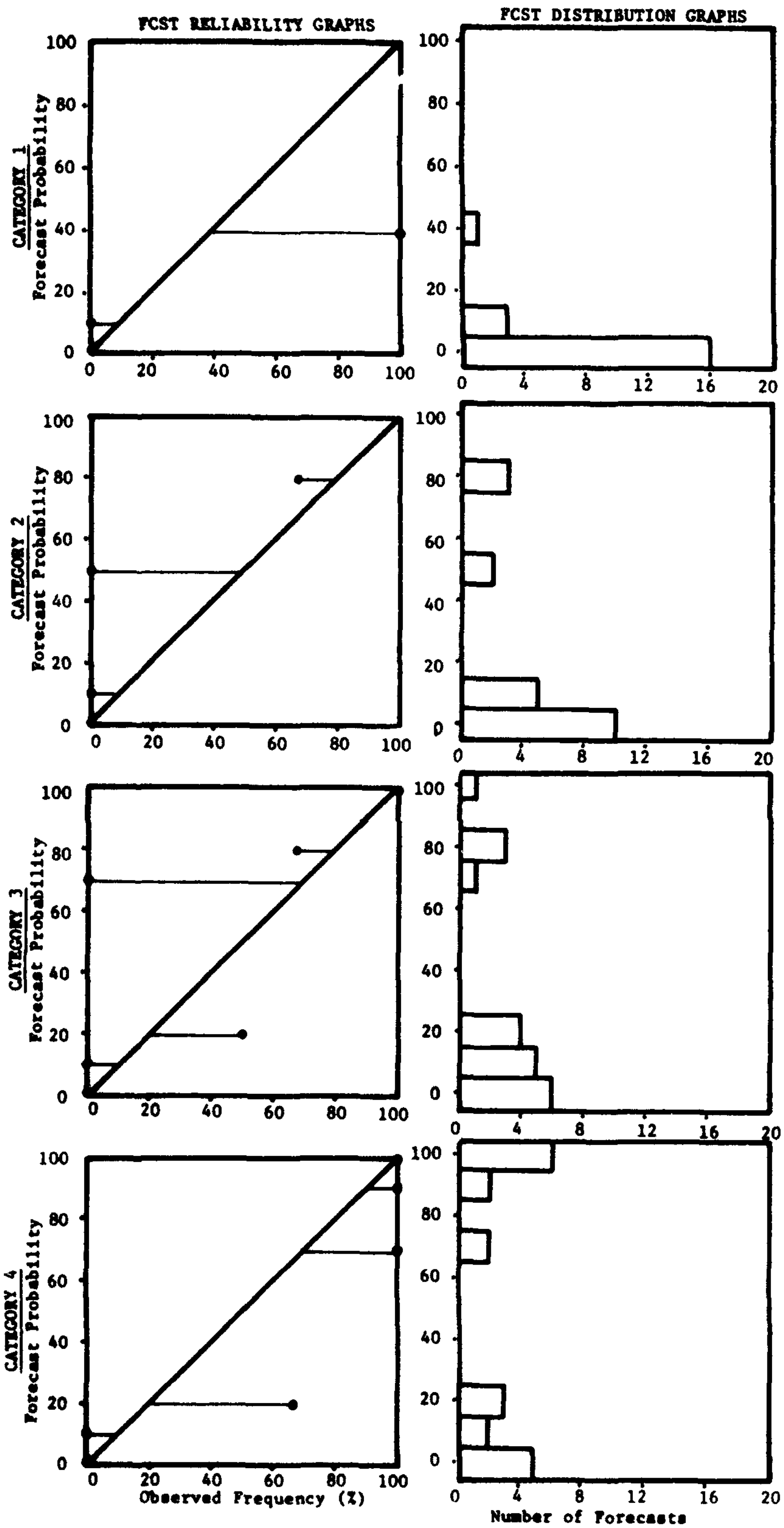


Figure 4-7. Example Forecast Distribution and Reliability Graphs for Four Category Evaluations.

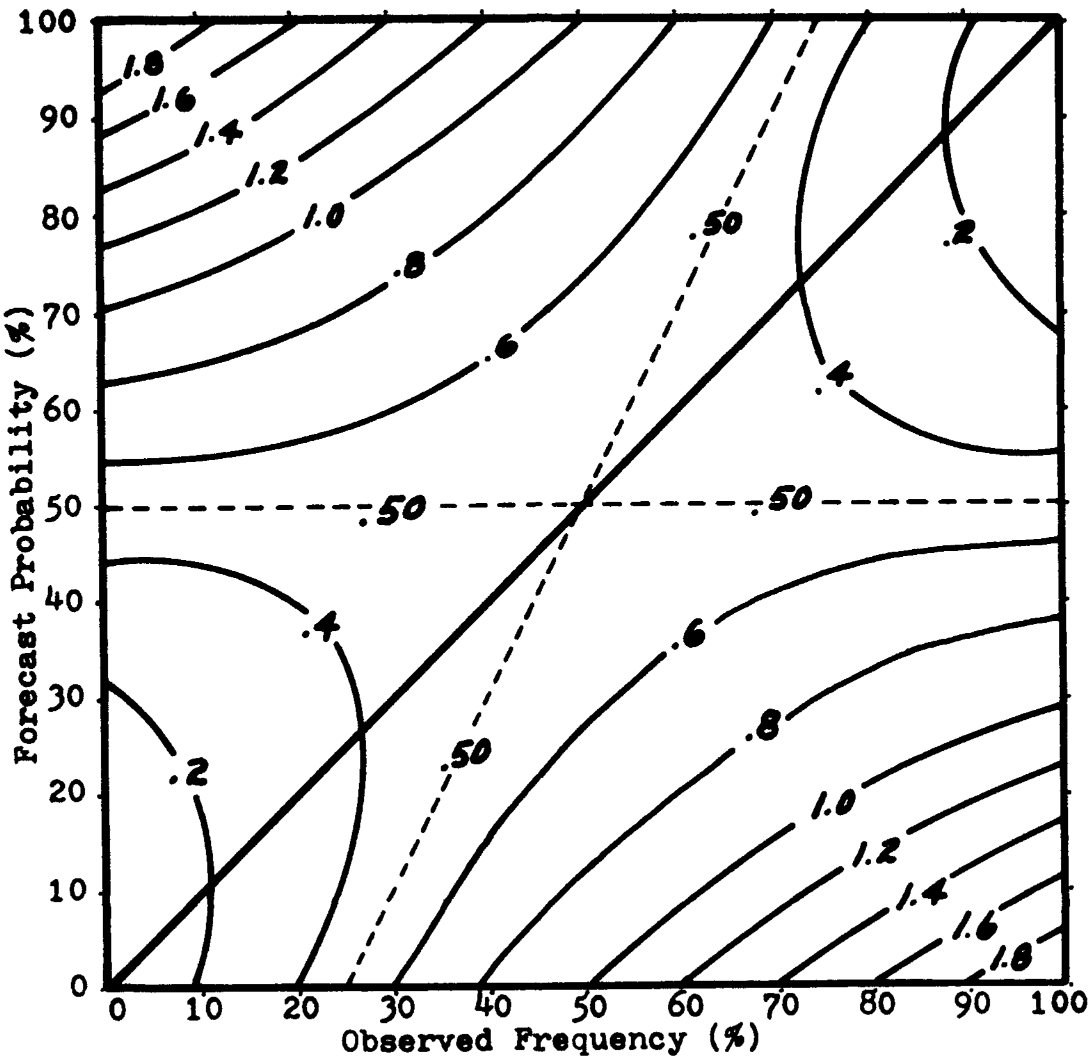


Figure 4-8. Brier Scores as a Function of Observed Frequency and Forecast Probability.

(2) The second example uses the forecasts in the first example and makes them perfectly sharp by lumping all probabilities above 50% in the 100% interval and those below 50% in the 0% probability interval. This result is a classic case of overconfidence. Note how the penalties are still reciprocal and how the Brier Score increases substantially only because of poor reliability.

(3) The third example demonstrates the combined effect of poor sharpness and poor reliability. Here, all the occurrences are evenly distributed in intervals above 50% probability, with nonoccurrences evenly distributed in intervals below 50%. This example illustrates the point discussed in paragraph 4-4b. The lowest (best) scores above 50% probability occur at 100% observed frequency, while below 50% probability they occur at zero observed frequency. Even though this gives a reasonably low Brier Score compared to the other two examples, the score would have been zero had the forecasts above 50% been assigned a probability of 100%, and those below 50% called 0% probability. This demonstrates how the Brier Score encourages sharpness. If skill permits, the best scores are attained when the extremes (0% or 100%) are used.

(4) The fourth example shows what the score would be if the forecasts had zero sharpness, i.e., a constant forecast probability equal to the sample climatological frequency. Such forecasts represent zero skill, but are perfectly reliable if used over a lengthy period.

Table 4-6. Example Brier Scores for Various Combinations of Sharpness and Reliability.

BRIER SCORE FOR POOR SHARPNESS AND PERFECT RELIABILITY							
PROBA-BILITY	# OF FCSTS	OCCURRENCES		OBSVD FREQ	NONOCCURRENCES		TOTAL PENALTY
		NUMBER	PENALTY		NUMBER	PENALTY	
1.0	10	10	.00	1.0	0	.00	.0
.9	10	9	.09	.9	1	.81	.9
.8	10	8	.32	.8	2	1.28	1.6
.7	10	7	.63	.7	3	1.47	2.1
.6	10	6	.96	.6	4	1.44	2.4
.4	10	4	1.44	.4	6	.96	2.4
.3	10	3	1.47	.3	7	.63	2.1
.2	10	2	1.28	.2	8	.32	1.6
.1	10	1	.81	.1	9	.09	.9
.0	10	0	.00	.0	10	.00	.0
TOTAL	100	50	7.00	.5	50	7.00	14.0
$PS = \frac{2 \times 14}{100} = 0.28$							
BRIER SCORE FOR PERFECT SHARPNESS AND POOR RELIABILITY							
1.0	50	40	0	.8	10	10	10
.0	50	10	10	.2	40	0	10
TOTAL	100	50	10	.5	50	10	20
$PS = \frac{2 \times 20}{100} = 0.40$							
BRIER SCORE FOR POOR SHARPNESS AND POOR RELIABILITY							
1.0	10	10	.0	1.0	0	.0	.0
.9	10	10	.1	1.0	0	.0	.1
.8	10	10	.4	1.0	0	.0	.4
.7	10	10	.9	1.0	0	.0	.9
.6	10	10	1.6	1.0	0	.0	1.6
.4	10	0	.0	.0	10	1.6	1.6
.3	10	0	.0	.0	10	.9	.9
.2	10	0	.0	.0	10	.4	.4
.1	10	0	.0	.0	10	.1	.1
.0	10	0	.0	.0	10	.0	.0
TOTAL	100	50	3.0	.5	50	3.0	6.0
$PS = \frac{2 \times 6}{100} = 0.12$							
BRIER SCORE FOR ZERO SHARPNESS AND PERFECT RELIABILITY							
.5	100	50	12.5	.5	50	12.5	25
$PS = \frac{2 \times 25}{100} = .50$							
BRIER SCORE FOR ZERO RELIABILITY AND PERFECT SHARPNESS							
1.0	50	0	.0	.0	50	50.0	50.0
.0	50	50	50.0	100.0	0	.0	50.0
TOTAL	100	50	50.0	.5	50	50.0	100.0
$PS = \frac{2 \times 100}{100} = 2.00$							

Table 4-7. Effect of Hedging on Brier Scores.

	PROBABILITY	# OF FCSTS	OCCURRENCES NUMBER	OCCURRENCES PENALTY	OBSVD FREQ	NONOCCURRENCES NUMBER	NONOCCURRENCES PENALTY	TOTAL PENALTY
INITIAL	1.0	5	5	.00	1.0	0	.00	.00
	.5	5	2	.50	.4	3	.75	1.25
	TOTAL	10	7	.50	.7	3	.75	1.25
		PS = 2 X 1.25/10 = .250						
ARTIFICIAL	1.0	5	5	.00	1.00	0	.00	.00
	.5	6	3	.75	.50	3	.75	1.50
	TOTAL	11	8	.75	.73	3	.75	1.50
		PS = 2 X 1.5/11 = .273						
HONEST	1.0	6	6	.00	1.0	0	.00	.00
	.5	5	2	.50	.4	3	.75	1.25
	TOTAL	11	8	.50	.7	3	.75	1.25
		PS = 2 X 1.25/11 = .227						
INITIAL	1.0	10	10	.00	1.00	0	.00	.00
	.9	9	8	.08	.89	1	.81	.89
	TOTAL	19	18	.08	.95	1	.81	.89
		PS = 2 X .89/19 = .094						
ARTIFICIAL	1.0	10	10	.00	1.00	0	.00	.00
	.9	10	9	.09	.90	1	.81	.90
	TOTAL	20	19	.09	.95	1	.81	.90
		PS = 2 X .9/20 = .090						
HONEST	1.0	11	11	.00	1.00	0	.00	.00
	.9	9	8	.08	.89	1	.81	.89
	TOTAL	20	19	.08	.95	1	.81	.89
		PS = 2 X .89/20 = .089						

(5) The last example depicts the opposite effect. To have zero reliability, all forecasts must be perfectly wrong, i.e., the event never occurs when the forecast probability is 100% and the event always occurs when the probability is 0%. Since the forecasts must be perfectly wrong, only 0 and 100% probabilities are possible. Thus, the forecasts are also perfectly sharp. Note that this is the only possible combination where the Brier Score reaches its maximum (2).

c. Hedging. A unique feature of the Brier Score is that it is a strictly proper scoring rule, i.e., a forecaster can maximize the expected score only by being completely honest in assigning probability values (Murphy, 1976b). This means that the Brier Score penalizes forecasters who try to "artificially" improve the reliability of their forecasts. Artifical improvement might be attempted, for example, if a forecaster has a particular interval in which the bias is positive (overforecasting). The reliability of that interval can be improved by calling a "sure case" (100% honest probability) a lower probability value equal to that of the unreliable interval.

(1) Table 4-7 illustrates two such cases. In each example, the initial verification represents the situation just prior to a hedging attempt. The "artificial" group illustrates the effect of placing the "sure" occurrence in the unreliable interval. The "honest" group shows the results that would be obtained, if the "sure" occurrence were properly placed in the 100% interval. In both instances, the "honest" assessment yields the better Brier Score. Although the penalty for improved reliability will decrease or disappear if hedging is used, the penalty for degraded sharpness is greater and produces a net increase in the score. Therefore, the only way to minimize the Brier Score is to make the forecasts just as good as skill allows, i.e., as high as possible when the event occurs and as low as possible when the event does not occur (Hughes, 1965).

d. Dependence of the Brier Score on Climatology and Number of Forecast Categories. Brier Score varies with the number of forecast categories and with the climatological frequency. These effects are shown below.

(1) The effect from the number of forecast categories on the Brier Score is demonstrated by the following example. Assume an equal climatological probability of the event occurring in each of the categories, i.e., for a two category system the event occurs 50% of the time in both categories, for a three category system the climatological probability is 33% for each category, etc. The general Brier Score equation can be modified and a zero skill Brier Score (PS_{zs}) computed for any number of categories (K) involved.

$$PS_{zs} = 1 - \frac{1}{K} \qquad (4\text{-}5)$$

Computed zero skill Brier Scores for a selected numnber of forecast categories are shown in Table 4-8.

Table 4-8. Variation of Brier Scores for Zero Skill and Number of Categories.

NO OF CATEGORIES	2	3	4	5	6	...	∞
CLIMATIC FREQ FOR EACH CATEGORY	50%	33-1/3%	25%	20%	16-2/3%	...	1/∞%
BRIER SCORE(PS_{zs})	.50	.67	.75	.80	.83	...	

Scores larger than these values indicate negative skill, while lower values represent positive skill. Remember the assumption made in calculating these zero skill scores when attempting to apply them. An equal distribution of climatic probabilities in each category is not very common; and if the distribution is unequal, the score for zero skill will change considerably. The significance of Table 4-8 is that, for a given skill level, one should expect Brier Scores for forecasts with a small number of categories to be lower than scores for forecasts with a larger number of categories.

(2) The climatological effect on the Brier Score can be seen intuitively by recalling that the score is the average of the squares of the differences between forecast and observed probabilities (paragraph 4-4a). For an extremely rare event, zero or very low forecast probabilities will be the general rule for any reasonable range of skill (positive or negative). Likewise, most of the observed frequencies will be zero. Consequently, the differences betwen the two probabilities will usually be small, and when squared and averaged, the resultant score will be even smaller (ref Category 1 in Table 4-4). The same reasoning applies to very frequent events except that both probabilities are very high, with very small differences. Thus, acceptable Brier Scores for events with very low or very high climatic frequencies will be much lower than for events with a frequency of 50%. Conversely, a large Brier Score (near 2) would result only if a large number of high (low) probabilities were forecast for rare (very frequent) events.

(3) The relationship between Brier Scores, climatology, and correlation of forecasts and observations for a two category system is depicted in Table 4-9. Correlation, as used here, is an approximation of forecasting skill where 0.99 reflects very high skill and 0.0 reflects zero skill. Note the small total variation in Brier Scores going from high skill to zero skill for an event with a climatic frequency of 1%, as opposed to the corresponding large variation for an event with a frequency of 50%. For the 1% event, 94% of the change in Brier Score occurs in the correlation range of 0.6 to 0.99; for the 50% event, 73% of the change is in the same range. This is significant, because that is usually the range of our forecasting skill. Now compare the maximum and minimum scores for various climatic frequencies. For example, the worst score for a 1% event is equal to the best score for an event with a climatic frequency of about 7% (interpolating). Hence, one must know the climatic frequency of the event before making judgments of forecasting skill. One can compute expected Brier Scores for events with climatic frequencies greater than 50% by using the complementary probability of the values given in Table 4-9. Similar tables for greater than two category forecasts are very complex due to the large number of possible combinations of frequencies and correlation.

e. Climatological Brier Scores. One cannot use the Brier Score above to interpret forecasting skill. The minimum score (0.0) represents perfect forecasts (positive skill); the maximum score (2.0) results from forecasts that are perfectly wrong (negative skill), i.e., all 0% forecasts verify with 100% observed frequencies and all 100% forecasts verify with 0% observed frequencies. Problems in interpreting Brier Scores arise because we do not know the value of the score for zero skill (somewhere between 0.0 and 2.0). A score for zero skill should be used to judge forecast performance. Several suggestions for controls (zero skill forecasts) with which to compare forecast performance are long-term climatology, sample climatology, conditional climatology, and TDL MOS forecasts. Methods for computing Brier Scores for these controls are shown below.

(1) If $C_1, C_2, C_3, \ldots, C_k$ are the respective climatological probabilities for categoreis 1,2,3,..., k, then, in the absence of any forecasting skill, the best values to choose for the forecast probability (R_{ij}) in the Brier Score equation will be the long term climatological probability (C_j) for all forecasts. This will minimize the Brier Score over the long-term and allows one to calculate a zero skill or climatological Brier Score (PS (C)) as follows (Panofsky and Brier, 1965):

$$PS(C) = 1 - \sum_{j=1}^{K} C_j^2 \qquad (4\text{-}6)$$

The above equation gives the climatological Brier Score for all categories combined. If climatological Brier Scores for individual categories are desired, they would be calculated by using the following relationship (Hughes, 1965):

$$PS(C)_j = C_j - C_j^2 \qquad (4\text{-}7)$$

As with regular Brier Scores, the sum of the scores for individual categories equals the overall climatological Brier Score:

$$PS(C) = \sum_{j=1}^{K} PS(C)_j \qquad (4\text{-}8)$$

Equations 4-7 and 4-8 provide an alternate method for computing overall climatological Brier Scores.

(a) Brier Scores for selected frequencies in forecasts with two categories were shown in the column for zero correlation (skill) of Table 4-9. To illustrate the computational procedures for any number of categories, assume that the long-term climatological frequencies for the four category verification example given earlier in Table 4-5 are as follows;
Category 1 - 2%, Category 2 - 12%, Category 3 - 21%, and Category 4 - 65%. Substituting these values in Equation 4-6, we obtain an overall climatological Brier Score as follows:

$$PS(C) = 1 - ((.02)^2 + (.12)^2 + (.21)^2 + (.65)^2) = .519$$

TABLE 4-9. EXPECTED BRIER SCORE AS A FUNCTION OF FORECAST CORRELATION AND CLIMATOLOGICAL FREQUENCY FOR A TWO CATEGORY SYSTEM.

CLIMATOLOGICAL FREQUENCY	CORRELATION 0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.95	.99
.01	.020	.020	.020	.020	.020	.019	.019	.018	.016	.012	.009	.004
.05	.095	.094	.094	.093	.091	.088	.083	.076	.066	.049	.036	.016
.10	.180	.179	.177	.174	.169	.161	.151	.136	.116	.085	.062	.028
.15	.255	.254	.251	.245	.236	.224	.208	.186	.157	.114	.082	.037
.20	.320	.319	.314	.306	.294	.277	.256	.228	.191	.138	.099	.045
.25	.375	.373	.367	.357	.342	.321	.295	.262	.219	.158	.112	.051
.30	.420	.417	.410	.398	.380	.357	.327	.290	.241	.173	.123	.055
.35	.455	.452	.444	.430	.410	.385	.352	.310	.258	.185	.132	.059
.40	.480	.477	.468	.453	.432	.404	.369	.325	.270	.193	.137	.062
.45	.495	.492	.482	.467	.445	.416	.379	.334	.277	.198	.141	.063
.50	.500	.497	.487	.471	.449	.420	.383	.337	.279	.200	.142	.064

NOTE: RELIABILITY CONTRIBUTION TO THE BRIER SCORE IS ASSUMED TO BE NEGLIGIBLE

Table 4-10. Climatological Brier Scores by Category Using Long-Term Climatology Compared with Actual Scores (from Table 4-5).

CATEGORY	COMPUTATIONS FOR $PS(C)_j$	ACTUAL SCORES (PS_j)
1	$PS(C)_1 = .02 - (.02)^2 = .020$	.020
2	$PS(C)_2 = .12 - (.12)^2 = .106$	.064
3	$PS(C)_3 = .21 - (.21)^2 = .166$	.131
4	$PS(C)_4 = .65 - (.65)^2 = .227$	.077
OVERALL	Using Eqn 4-8, PS(C) = .519	.292

By using equation 4-8, corresponding scores for individual categories can be calculated as in Table 4-10:

These scores indicate that the forecasts exhibited positive skill overall compared to climatology, because the overall actual Brier Score (.292) was lower than the climatological Brier Score (.519). Zero skill existed in Category 1 and positive skill is evident in the others; i.e., actual scores are lower than the climatological scores.

(2) Difficulties may arise from using long-term climatology as a control, because the observed frequency of the event for the evaluation period generally will be different from long-term climatology. Another approach is to use sample climatology as the control, i.e., the observed frequency of the event in the evaluation period. This may not represent a true zero skill, because the sample climatology would not be known prior to issuing the forecasts (Hughes, 1965; Glahn and Jorgensen, 1970). However, when long-term climatology is not available, a Brier Score based on sample climatology may be the best control.

(3) Another method for evaluating the quality of a set of forecasts is to compare Brier Scores with forecasts for the same event which have been produced by other means. Brier Scores for conditional climatology forecasts can be computed by using the procedures described for ordinary forecasts (paragraph 4-4a). These scores could then be used to determine if actual skill was better than the skill of conditional climatology. Similar comparisions could be made for any other like forecast, e.g., TDL MOS, NWS probability of precipitation, etc.

f. Ratio Skill Score. A measure frequently used to evaluate the skill in a set of probability forecasts is the ratio skill score. This score is the percentage improvement of the forecasts being evaluated over a control which is assumed to represent zero skill. It ranges from 100% for a perfect score (PS=0) to minus infinity. Compared to the control, scores above zero indicate positive skill; a score of zero indicates no skill; scores below zero indicate negative skill.

(1) The ratio skill score (RSS) used to evaluate the Brier Score (PS) for a set of probability forecasts against the Brier Score (PS (C)) for long-term climatology is computed by (Hughes, 1967a)

$$RSS(PS(C)) = \left\{\frac{PS(C) - PS}{PS(C)}\right\} 100\% \quad (4\text{-}9a)$$

or $$RSS(PS(C)) = (1 - PS/PS(C))\ 100\% \quad (4\text{-}9b)$$

Table 4-11 shows the ratio skill scores for the scores in Table 4-9.

(2) Ratio skill score can be computed by comparing any two sets of forecasts; e.g., man-made forecasts, long-term climatology, sample climatology, conditional climatology, and TDL MOS forecasts. Enter the Brier Scores for the two forecast systems being compared into either equation 4-9a or 4-9b.

(3) If conditional climatology is available for the event, the ratio skill score for conditional climatology would be a good baseline for determing the quality of a set of forecasts. The main advantage is that the sample climatology is the same in both forecasts; thus, the problems discussed in paragraph 4-4b are eliminated.

Table 4-11. Ratio Skill Scores (RSS/PS(C)) Corresponding to Expected Brier Scores for Forecasts with Two Categories Shown in Table 4-9. (Multiply by 100 to obtain percentages.)

CLIMO %	CORRELATION 0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.95	.99
1	.0	.0	.0	.0	.0	.05	.05	.10	.20	.40	.55	.80
5	.0	.01	.01	.02	.04	.07	.13	.20	.31	.48	.62	.83
10	.0	.01	.02	.03	.06	.11	.16	.24	.36	.53	.66	.84
15	.0	.00	.02	.04	.07	.12	.18	.27	.38	.55	.67	.85
20	.0	.00	.02	.04	.08	.13	.20	.29	.40	.57	.69	.86
25	.0	.01	.02	.05	.09	.14	.21	.30	.42	.58	.70	.86
30	.0	.01	.02	.05	.10	.15	.22	.31	.43	.59	.71	.87
35	.0	.01	.02	.05	.10	.15	.23	.32	.43	.59	.71	.87
40	.0	.01	.03	.06	.10	.16	.23	.32	.43	.60	.71	.87
45	.0	.01	.03	.06	.10	.16	.23	.33	.44	.60	.71	.87
50	.0	.01	.03	.06	.10	.16	.23	.33	.44	.60	.72	.87

4-5. Summary. This chapter discussed two methods for evaluating probability forecasts: sharpness and reliability measures and the Brier Score.

a. Sharpness and reliability are evaluated either by inspecting the verification statistics or by plotting graphs. Detailed analyses permit the identification of specific biases and provide clues for correcting deviations from acceptable sharpness and reliability. Interpretation of skill is simpler than with the Brier Score. A disadvantage is that sharpness and reliability measures do not provide a single number measure of goodness; this makes it difficult to assess forecast trends.

b. Since the Brier Score does not indicate skill directly, it must be compared with the score of some control, such as climatology of conditional climatology, to obtain a measure of performance. If the ratio skill score is used for the comparison, the single number result makes it easy to evaluate forecast trends. However, interpretation of the score is not simple, and comparisons of scores must be made with caution. Other disadvantages are that it requires a substantial amount of computations and only indicates overall performance.

c. Since both methods fulfill different needs, the optimum evaluation effort would use both techniques. The Brier Score indicates overall performance; sharpness and reliability measures identify specific forecast problems. If only one evaluation method is used, the choice is to compute sharpness and reliability measures.

Chapter 5

PROBABILITIES IN DECISION MAKING

5-1. Introduction. Although customers must make their own decisions, forecasters and SWOs must also be knowledgeable of their decision problems to properly integrate weather support. We are the weather experts. Recipients of our support are generally not well versed in the use of information we can furnish (especially in probabilistic form). Thus, we have an inherent responsibility to furnish the guidance needed to use our forecasts most effectively (Glahn, 1964). Since decision theory is a complete field of study in itself, this section will only introduce some of the simpler techniques which can be applied to weather-related decision problems. Specifically, it describes a general decision matrix, illustrates applications of the simple cost-loss model, defines critical probability, and demonstrates methods for calculating the value of forecast information.

5-2. General Decision Matrix. Many schemes are used to aid the decision maker. Some of these apply to situations where the outcome is known with complete certainty. Others are effective in situations where we know nothing about the outcome. Finally, some apply to situations where we have only partial knowledge of future events. The first of these situations does not concern us; nor should the second. The third situation is decision making under risk, and considers that one of two or more future events may occur, each with a specified probability. We can apply this last case to meteorological situations in which the frequencies of the various future weather states are estimated or predicted, i.e., probability forecasts (Epstein, 1962). A matrix is the most convenient method for summarizing all the elements involved in weather decision problems. The generalized form of a decision matrix which uses expenses as a measure of value is shown in Table 5-1. It can be used directly, or serve as the framework for developing specialized models.

Table 5-1. General Expense Matrix (Murphy, 1976b).

ACTIONS	STATES OF WEATHER					EXPECTED EXPENSE (E)
	W_1	...	W_n	...	W_N	
a_1	e_{11}	...	e_{1n}	...	e_{1N}	$E_1 = \sum_{n=1}^{N} P_n e_{1n}$
⋮	⋮		⋮		⋮	....
a_m	e_{m1}	...	e_{mn}	...	e_{mN}	$E_m = \sum_{n=1}^{N} P_n e_{mn}$
⋮	⋮		⋮		⋮	....
a_M	e_{M1}	...	e_{Mn}	...	e_{MN}	$E_M = \sum_{n=1}^{N} P_n e_{Mn}$
PROBA-BILITY	P_1	...	P_n	...	P_N	

a. Explanation.

(1) In the general matrix, all possible courses of action, strategies, or decision options under consideration are listed in the left column, i.e., $a_1 \ldots a_m \ldots a_M (m=1,2,\ldots,M)$. Under the states of weather, the notations, $W_1 \ldots W_n \ldots W_N \ldots$ $(n=1,2,\ldots,N)$, represent the various weather thresholds which affect one or more courses of action. For each action-state pair (a_m, W_n), there is a corresponding consequence or outcome (e_{mn}), which

represents the expense for that course of action if that state of weather occurs (Murphy, 1966). For example, if action, a_1, is implemented and if weather state, W_1, occurs, the associated expense is e_{11}.

(2) Each weather probability, P_n, corresponds to a particular weather threshold or state of weather, W_n. P_n represents the probability that the state of weather, W_n, will occur. Additionally, the sum of all the weather probabilities must equal one ($P_1 + ... + P_n + ... + P_N = 1$).

(3) Given the expense associated with each action-state pair and the probabilities of each state of weather, the long-term expected expense (E) can be calculated by using the equations in the right hand column. The expected expense is simply the weighted average of the expenses associated with each action-state pair, where the weights are the corresponding weather probabilities. For example, the expected expense for action, a_1, would be computed as follows:

$$E_1 = P_1 e_{11} + ... + P_n e_{1n} + ... + P_N e_{1N} \quad (5\text{-}1)$$

If the decision maker wants to minimize expenses (losses), his course of action is the one which yields the smallest value for E_m (Murphy, 1976b), i.e., that course of action which will cost the decision maker the least amount over the long-term, provided that the probabilities are reliable.

b. Example. Consider the situation in which a wing commander must decide between four ways to protect his aircraft, when threatened by winds.

(1) Table 5-2 sets up the decision problem in matrix form. Four wind thresholds are listed under the states of weather. The model can help decide which action to take, regardless of the cause of the threat. If the wing commander wants to minimize expected costs, the costs associated with each consequence (e_{mn}) must be obtained and entered in the matrix. For example, we will consider only two types of costs.

Table 5-2. Incomplete Cost Matrix for Protection Against Wind Damage.

ACTIONS	STATES OF WEATHER				EXPECTED COSTS (E)
	W_1 = WIND <30 kts	W_2 = WIND ≥30<50 kts	W_3 = WIND ≥50<65 kts	W_4 = WIND ≥65 kts	
a_1 = No Protection					
a_2 = Tie Down					
a_3 = Hangar					
a_4 = Evacuate					
PROBABILITY	P_1 =	P_2 =	P_3 =	P_4 =	

(a) First, is the cost of taking each of the actions indicated. Assume that the figures given in Table 5-3 reflect the costs obtained from the customer. They include such factors as manpower required to tie down, hangar, and unhangar aircraft; and for evacuation, flying costs to and from the refuge base, TDY expenses, and non-routine costs generated by action taken.

Table 5-3. Costs of Taking Protective Action (Thousands of Dollars)

ACTION	W_1	W_2	W_3	W_4
a_1	$ 0	$ 0	$ 0	$ 0
a_2	1	1	1	1
a_3	4	4	4	4
a_4	120	120	120	120

(b) The other costs would be the estimated costs or losses as a result of damage sustained when the aircraft are not protected or when the protection is inadequate. Table 5-4 represents these costs. These figures would also be supplied by the customer. [1]

1AWSP 178-2 provides guidance in computing cost figures. Standard cost factors are included in AFR 173-10, Vol I.

Table 5-4. Potential Losses Due to Wind Damage (Thousands of Dollars)

ACTION	W_1	W_2	W_3	W_4
a_1	$ 0	$300	$1500	$12000
a_2	0	0	600	6000
a_3	0	0	0	1500
a_4	0	0	0	0

The potential loss varies with the degree of protective action taken and the severity of wind thresholds.

(c) To obtain the total costs or expenses associated with each consequence (e_{mn}) of the decision matrix, the corresponding values in Tables 5-3 and 5-4 must be added. Table 5-5 shows the resultant matrix. It is now ready to apply to a decision problem.

Table 5-5. Cost Matrix for Protection Against Wind Damage Prior to Use (Thousands of Dollars)

ACTION	STATES OF WEATHER				EXPECTED COSTS (E)
	W_1 <30 kts	W_2 ≥30 & <50 kts	W_3 ≥50 & <65 kts	W_4 ≥65 kts	
a_1 = No Protection	$ 0	$300	$1500	$12,000	E_1 =
a_2 = Tie Down	1	1	601	6,001	E_2 =
a_3 = Hangar	4	4	4	1,504	E_3 =
a_4 = Evacuate	120	120	120	120	E_4 =
PROBABILITY	P_1 =	P_2 =	P_3 =	P_4 =	

(d) Assume a hurricane threatens the installation, and the forecast probabilities for the different states of weather 12 hours from now are as follows: $P(W_1) = 5\%$, $P(W_2) = 80\%$, $P(W_3) = 10\%$, and $P(W_4) = 5\%$. Expected costs (E_m) are shown below:

$$E_m = P_1 e_{m_1} + P_2 e_{m_2} + P_3 e_{m_3} + P_4 e_{m_4} \qquad (5\text{-}2)$$

$$E_1 = .05 \times 0 + .8 \times 300 + .1 \times 1500 + .05 \times 12{,}000 = \$\,990$$

$$E_2 = .05 \times 1 + .8 \times 1 + .1 \times 601 + .05 \times 6{,}001 = \$\,361$$

$$E_3 = .05 \times 4 + .8 \times 4 + .1 \times 4 + .05 \times 1{,}504 = \$\,79$$

$$E_4 = .05 \times 120 + .8 \times 120 + .1 \times 120 + .05 \times 120 = \$\,120$$

After entering the probabilities and expected costs in their appropriate matrix positions, we obtain the final decision matrix in Table 5-6.

Table 5-6. Final Cost Matrix for Protection Against Wind Damage (Thousands of Dollars)

ACTIONS	STATES OF WEATHER				EXPECTED COSTS (E)
	W_1 <30 kts	W_2 ≥30 & <50 kts	W_3 ≥50 & <65 kts	W_4 ≥65 kts	
a_1 = No Protection	$ 0	$ 300	$ 1,500	$12,000	$ 990
a_2 = Tie Down	1	1	601	6,001	361
a_3 = Hangar	4	4	4	1,504	79
a_4 = Evaluate	120	120	120	120	120
PROBABILITY	P_1 = .05	P_2 = .80	P_3 = .1	P_4 = .05	

(e) The decision rule assumed earlier is that the preferred choice is the course of action which results in the least expected cost. Thus, action a_3 (hangar the aircraft) is preferred for this set of probabilities. Various combinations of probabilities yield different values of expected costs, and, thus, different decisions. However, when one course of action affords total protection, such as evacuate (a_4), the expected cost (E) of that action remains unchanged.

(2) Two key assumptions in this decision process are that the probabilities are reliable, and that the expected costs are long-term averages. The effect of the latter assumption is shown by one of the computations for expected costs. Computation of E_1 shown under equation 5-2 above is repeated for illustration:

$$E_1 = .05 \times 0 + .8 \times 300 + .1 \times 1500 + .05 \times 12{,}000 = \$990 \qquad (5\text{-}3)$$

The first component of E_1 contributes nothing to the expected cost, because there is no potential loss (i.e., no damage will occur as long as the winds are less than 30 knots). In the second component, the .8 means that 8 times out of 10 (reliable forecasts assumed) the winds will be within that threshold (≥ 30& < 50 kts). On each of those eight occasions the damage will amount to $300K with no damage on the other two days (total - $2400K). The average damage amount is $2400K divided by 10 occasions or $240K which is .8 X 300. Similar reasoning applies to the remaining components. Thus, if the forecasts are totally reliable (bias = 0), average costs will equal the expected costs in the long-term. Otherwise, actual costs will differ in proportion to the net reliability error (bias).

(3) Notice that sharpness is not the main issue here. Intelligent decisions can still be made without a large degree of sharpness. As long as the probabilities are reliable and do not cluster around the climatic frequency, they are useful in decision making. However, credibility is soon lost, if discrimination between events (high and low probabilities) does not approximate the state-of-the-art. The effect of reliability and sharpness on expected and actual costs will be addressed later.

5-3. Utilities.

a. Background. Money (dollar value) is the most common unit of value used to represent consequences of decision actions (e_{mn}). However, as a unit of value, money has one very serious deficiency. Since a decision matrix is a model of the thought process of the decision maker, monetary value frequently does not adequately represent the importance that a decision maker assigns to the consequences. Further, it is very difficult to assign a monetary value to many types of consequences such as loss of military readiness, political impact, loss of prestige, loss of human life, and reduced combat effectiveness. Thus, non-monetary considerations may, and frequently do, influence the value a decision maker places on particular outcomes (Murphy, 1976b).

b. Utility. The term "utility" is used as the unit of value of consequences, when non-monetary factors are involved. Utility is an all encompassing term which reflects a decision maker's true value (preference or importance weight) associated with a given consequence or outcome (Murphy, 1976b). Utilities combine monetary factors such as costs, losses, or profits with non-monetary factors like opportunity loss, risk, or desirability, to form a dimensionless number which represents the true value of the consequence to a decision maker. Thus, different decision makers may have different utilities, and an individual's utilities may change, as factors which influence the decisions vary.

c. Utility Matrices. A utility matrix takes the same form as the general expense matrix (Table 5-1). The only difference is substituting utility value (e_{mn}) for expenses (emn) for each consequence, and expected utility (U) for expected expense (E). Utility values are either positive or negative. The objective is to maximize positive utilies, such as profits or economic gain, and to minimize negative utilities.

d. Transformation of an Expense Matrix into an Equivalent Utility Matrix. There are a number of ways to determine a customer's utilities. A formal method, in terms of regret, is given in Attachment 6. Other approaches will be described later. In general, if a customer's utilities are linearly related to the respective expenses of the consequences, an expense matrix can be transformed directly into an equivalent utility matrix with an arbitrary scale ranging from 0 to 1. The equation for performing the transformation is given by (Murphy, 1976a):

$$u_{mn} = (e_{mn} - e_L)/(e_M - e_L) \qquad (5\text{-}4)$$

where:

u_{mn} = the utility value equivalent to expense e_{mn} (ranges 0 to 1).

e_{mn} = the expense (value) of the consequence being transformed.

e_L = expense value of the least preferred consequence.

e_M = expense value of the most preferred consequence.

By using this transformation, the **most preferred** consequence, e_M (the one of **least expense**) takes on the utility value, $u_{mn} = 1$ (the **greatest utility**). Likewise, the **least preferred** consequences, e_L **(the largest expense)**, transforms to the utility value, $u_{mn} = 0$ (the least utility).

(1) Example. Table 5-7a is an abbreviated form of the general expense matrix shown in Table 5-1. We will use this table to demonstrate the transformation technique described above.

Table 5-7a. Abbreviated Expense Matrix.

$e_{11} = -5$	$e_{12} = 70$	$e_{13} = 85$
$e_{21} = 5$	$e_{22} = 50$	$e_{23} = 90$
$e_{31} = 15$	$e_{32} = 30$	$e_{33} = 95$

Table 5-7b. Abbreviated Equivalent Utility Matrix.

$u_{11} = 1.0$	$u_{12} = .25$	$U_{13} = .10$
$u_{21} = .9$	$u_{22} = .45$	$u_{23} = .05$
$u_{31} = .8$	$u_{32} = .65$	$u_{33} = .00$

From Table 5-7a we find that the most preferred consequence (e_M) is e_{11}, and the least desired (e_L) is e_{33}. In this example equation 5-4 takes the form:

$$u_{mn} = (e_{mn} - 95)/(-5-95) = (e_{mn} - 95)/-100 \qquad (5\text{-}5)$$

Substituting values for e_{mn}, we obtain the equivalent utility values u_{mn} shown in Table 5-7b.

(2) Such a transformation is useful for two reasons. First, it assigns the highest utility value (1) to the most preferred consequence, and places the decision objective of maximizing utilities in a positive sense. Second, it establishes a standard scale from 0 to 1 to which the customer can better relate by using ratios to confirm whether or not the equivalent utilities do in fact reflect true preferences. If adjustments to the utilities are required, this scale simplifies and expedites the modifications. In fact, all utility matrices should be checked before use to see if they reflect true preferences. If not, the equivalent utilities should be modified, or another approach used to develop true utilities.

5-4. Original Cost-Loss Model. The literature on probability forecasting frequently makes reference to the "cost-loss" model. The cost-loss model is a very simple and specialized case of the general decision model given earlier. It provides a realistic description of situations faced by many decision makers and is extensively used by meteorologists and others in the civilian community. This model was originally developed to describe a situation where a decision maker must decide whether or not to take protective action with respect to some activity or operation based on an uncertain forecast of adverse weather. However, it also has other applications when only two courses of action are under consideration. Following a format similar to the general matrix, the original cost-loss model is depicted in Table 5-8.

Table 5-8. Matrix for Original Cost-Loss Model

ACTIONS	STATES OF WEATHER		EXPECTED COST (E)
	Adverse	Not Adverse	
a_1 = Protect	Cost (C)	Cost (C)	$E_1=P_1C+P_2C = C$
a_2 = No Protection	Loss (L)	0	$E_2=P_1L$
PROBABILITY	P_1	P_2	

a. Terms. In this model, the cost of protection is denoted by C. It is assumed that, when protective action is taken, the resources are completely protected against adverse weather. Thus, the cost of the two consequences associated with the first course of action, a_1, are each equal to C. The loss which results when no protective action is taken and adverse weather occurs is denoted by L. Finally, no cost or loss results when no protection is taken and the weather is not adverse therefore, the cost is zero (Murphy 1976a).

b. Explanation. Expected costs are calculated as indicated. $E_1 = C$ since $P_1 + P_2 = 1$. Now assume the decision maker wants to select the action which minimizes expected costs. A simple decision rule for this situation is determined by equating the two expected costs (E_1 and E_2) and solving for the probability, P_1. Thus, when $P_1 = C/L$ (the cost-loss ratio), the expected costs are equal. On the other hand, if $P_1 > C/L$, the expected cost is least for action, a 1 (protect). However, for $P_1 < C/L$, action, a 2 (no protection), yields the least expected cost. This decision rule can be summarized as follows (Murphy, 1976a):

Protect (a_1) if $P_1 > C/L$

Indifferent (a_1 or a_2) if $P_1 = C/L$ (5-6)

No Protection (a_2) if $P_1 < C/L$

To make economic sense, the ratio, C/L, must have a total range between zero and unity. Consider, for example, the possibility that $C/L > 1$. In this case the cost of protection would exceed the loss and it would be uneconomical to protect against adverse weather at all. Similarly, negative values of C/L are economically meaningless (Thompson and Brier, 1955).

c. Example. Assume that the base civil engineers (BCE) finished pouring fresh concrete just before quitting time. If any measurable amount of rainfall occurs within the next 12 hours, they must refinish the surface at a cost (loss) of $3000 (materials, plus labor). However, a portable cover could be placed over the concrete, at a cost of $450 in overtime pay. The most economical course of action for this problem can be determined very quickly by computing the cost-loss ratio (C/L) and comparing it to the probability of measurable rainfall. For this situation, C/L = 450/3000 = .15. By using Eq 5-6, the concrete should be covered, if the probability of measurable rainfall (P_1) is greater than 15%.

5-5. General Cost-Loss Model. The basic cost-loss model assumes that protective action completely eliminates losses due to adverse weather. However, in many situations all resources cannot be protected; in others, the protective actions available to the decision maker may only reduce the losses. A more general version of the cost-loss model which accounts for unprotectable losses is shown in Table 5-9 (Murphy, 1976a).

Table 5-9. Matrix for the General Cost-Loss Model.

ACTION	STATES OF WEATHER		EXPECTED EXPENSES (E)
	ADVERSE	NOT ADVERSE	
a_1 = Protect	$C + \ell$	C	$E_1 = C + P_1\ell$
a_2 = No Protection	$L + \ell$	0	$E_2 = P_1(L + \ell)$
PROBABILITY	P_1	P_2	

a. Terminology. Terms are the same as in the basic model, except that unprotectable losses (ℓ) have been included. Thus, the total loss that could be incurred is L + ℓ.

b. Explanation. By using the logic applied in the original model, it follows that the expected costs are equal when $P_1 = C/L$; thus, the decision rules for both models are identical (ref equation 5-6). Consequently, from a decision making standpoint, only the **protectable** portion of the potential loss (L) needs to be specified in each model (Murphy, 1976a). For example, if this model were applied to a wind damage decision situation, one would not need to include such unprotectable items (ℓ) as buildings, fixed towers, fences, etc, unless they are afforded protection by the action taken. But windows that are covered, antennas or towers that are taken down, etc, in the threat of high winds would be included in the potential loss (L). Although this model offers no advantages over the other in decision making roles, it does show how expected costs would be computed when they are needed for value analysis, etc. Note that neither of these models provides a means for considering variable costs such as labor in a snow removal situation (Kernan, 1975).

5-6. Critical Probability. In the discussion of cost-loss models, we derived a decision rule in which the cost-loss ratio determined the probability threshold above which protective action should be taken. Critical probability as used in Air Weather Service is an extension of the cost-loss ratio concept, in that it can be applied to any two-by-two action-state decison matrix.

a. Derivation. Critical probability (Pc) may be derived using the procedure of the cost-loss ratio and given the consequences A, B, C, and D (in utility units) from Tables 5-10a & b below.

Table 5-10a. Protection Matrix for Definition of Critical Probability.

ACTION	STATES OF WEATHER		EXPECTED UTILITIES (U)
	Storm/Rain	No Storm/Rain	
a_1 = Protect	A	C	$U_1 = P_1A + P_2C$
a_2 = No Protection	B	D	$U_2 = P_1B + P_2D$
PROBABILITY	P_1	P_2	

Table 5-10b. Launch Matrix for Definition of Critical Probability

ACTION	STATES OF WEATHER		EXPECTED UTILITIES (U)
	Favorable	Unfavorable	
a_1 = Go	A	C	$U_1 = P_1A + P_2C$
a_2 = No Go	B	D	$U_2 = P_1B + P_2D$
PROBABILITY	P_1	P_2	

$$P_c = \frac{C-D}{B+C-A-D} \qquad (5\text{-}7)$$

The corresponding decision rule for a critical probability is:

$$\text{Act } (a_1) \text{ if } P_1 > P_c$$

$$\text{Indifferent } (a_1 \text{ or } a_2) \text{ if } P_1 = P_c \qquad (5\text{-}8)$$

$$\text{No Action } (a_2) \text{ if } P_1 < P_c$$

(1) Critical probability is the threshold or breakeven probability above which it is cost effective for a decision maker to take a specific action, i.e., the long-term positive utility (value, payoff, etc) is maximized and the negative utility (cost, loss expense, regret, etc) is minimized. It may be based on monetary value or other measures of utility. Note that the critical probability must be stated in terms of the weather event which causes the action to be taken. This is a subtle, but important, point and is the reason two different examples are given. In the first matrix, action is taken when unfavorable or adverse weather (storm, rain, etc) threatens; in the second case, the action is associated with favorable weather.

(2) Equation 5-7 reduces to Pc =C/L for the original cost-loss model (see Table 5-8) because A = C, B=L, and D =0 for the cost-loss model.

b. Matrix Example. Consider an airborne training operation as depicted in the matrix of Table 5-11.

Table 5-11. Airborne Training Matrix (Dollars)

ACTIONS	STATES OF WEATHER		EXPECTED EXPENSES (E)
	Favorable	Unfavorable	
a_1 = Fly	A = +1000	C = -5700	$E_1 = 1000P_1 - 5700P_2$
a_2 = Stand down	B = -1200	D = -1200	$E_2 = -1200P_1 - 1200P_2$
PROBABILITY	P_1	P_2	

(1) Definitions.

A is the benefit realized by the customer when the weather is favorable and the mission goes. In this case, the benefit less operating costs is $1,000.

B is the cost (negative benefit) incurred if the customer stands down and the weather is favorable - a lost opportunity. A missed training day costs $1200 in additional TDY funds.

C is the cost or loss if the customer takes action, but, because of unfavorable weather, cannot accomplish the mission (aborts). Each training mission is a three-hour flight by a C-141. If the mission is aborted because of unfavorable drop zone weather, the costs would be $4,500 (3 hrs X $1500/hr) plus $1200 for another TDY day (total = $5700).

D is a cost or benefit. If there is a cost for mission delay, then it is a cost. If a delay has no cost, then the abort cost can be saved and D is a cost avoidance benefit (correct stand down). The customer considers this a delay cost of $1200 in this example.

P_1 is the probability that no weather factors (ceiling, visibility, wind, hazards, etc) will cause mission cancellation, abort, or failure from take-off to recovery. This is called a tailored probability forecast. Recall that $P_1 + P_2 = 1$ and therefore, $P_2 = 1 - P_1$.

(2) Explanation. Applying equation 5-5, the critical probability for this example is:

$$P_c = \frac{-5700 + 1200}{-1200 - 5700 - 1000 + 1200} = .67 \qquad (5\text{-}9)$$

Thus, the decision rule (equation 5-6) for this decision problem is:

Fly if $P_1 > .67$

Indifferent if $P_1 = .67$ (5-10)

Standdown if $P_1 < .67$

Referring to the matrix in Table 5-11, this means that the expected expense (E) for each mission will equal $1200 when the probability of favorable weather, $P_1 = P_c = .67$. As P_1 increases, E decreases, and E_1 increases because of the weights exerted by the probabilities in the equation for expected expenses. The reverse occurs when P_1 decreases.

(3) Transformation to Utilities. In the example above, the critical probability of 67% results in a significant number of missedopportunities. Suppose the Army unit commander complained about the recent number of cancellations due to weather, and stated that it is essential for their airborne unit to complete 12 missions during the next 20 days. Also assume the squadron commander of the C-141 unit that supports the Army commander just received a notice that their fuel supplies and TDY funds are low and must be conserved. Faced with this situation, both commanders ask the SWO to help work out a compromise in the critical probability used for making their launch decisions.

a. Applying the utility transformation equation (5-3) to the expense matrix (Table 5-11), the SWO prepared an equivalent matrix (Table 5-12) and showed it to the squadron commander. The commander was appalled at the importance weight indicated by the utility value (B = .67) for a stand down with favorable weather (missed opportunity). He was satisfied with the most preferred (A = 1) and least preferred consequences (C = 0), but the other two did not reflect his true preferences in the present situation. After discussion between the SWO and the two commanders, consequence B was adjusted to a value of .1 because now this consequence was considered nearly as undesirable as the least preferred consequence. This action should significantly reduce the number of missed opportunities and satisfy the Army unit commander. Consequence D was also adjusted to a lower value (.6). This has the effect of slightly increasing the possibility of aborts, but the squadron commander reasoned that they could save fuel and TDY funds in the long run. The extra training missions they were flying could be reduced since the number of operational missions should increase.

Table 5-12. Airborne Training Equivalent Utility Matrix.

ACTIONS	STATES OF WEATHER		EXPECTED UTILITIES (U)
	Favorable	Unfavorable	
a_1 = Fly	A = 1	C = 0	$U_1 = P_1$
a_2 = Stand down	B = .67	D = .67	$U_2 = .67$
PROBABILITY	P_1	P_2	

Table 5-13. Modified Airborne Training Utility Matrix

ACTIONS	STATES OF WEATHER		EXPECTED UTILITIES (U)
	Favorable	Unfavorable	
a_1 = Fly	A = 1	C = 0	$U_1 = P_1$
a_2 = Stand down	B = .1	D = .6	$U_2 = .1P_1 + .6P_2$
PROBABILITY	P_1	P_2	

(b) With these adjustments in utilities (Table 5-13), a modified critical probability is calculated.

$$P_c = \frac{0 - .6}{.1 + 0 - 1.0 - .6} = .4 \qquad (5\text{-}11)$$

Therefore, the new decision rule becomes:

$$\begin{aligned} &\text{Fly if } P_1 > .4 \\ &\text{Indifferent if } P_1 = .4 \\ &\text{Standdown if } P_1 < .4 \end{aligned} \qquad (5\text{-}12)$$

The squadron commander states that he is much more comfortable with this rule because it reduces the number of lost opportunities and should help satisfy the Army unit commander's needs.

c. Operational Verification by Using Critical Probability (P_c). Another way to help a customer choose the proper critical probability value is to show the operational verification that would result when different probability values are used. By inspecting the number of hits, successes, false alarm aborts, missed opportunities, correct stand downs, etc, the customer can readily assess the effect different critical probability values would have on the operation. In fact, the customer should be provided this information for the P_c value chosen regardless of how it was selected.

(1) Preparation. Part A of Table 5-14 below shows the type of verification results that would normally be prepared for any set of probability forecasts. Suppose these were in support of photo reconnaissance operations where the weather threshold over the target was 3/8 or less cloud cover below 10,000 feet for a specific period. To build a series of matrices showing the number of successful launches, missed opportunities, and aborts that could be expected for various critical probabilities, you need to know the number of forecasts and event occurrences that would have resulted from using the different probabilities. Part B or Table 5-14 shows these distributions. They were obtained by cumulative summation of numbers below, and equal to or above, the critical probability value. Individual verification matrices were then prepared using these values as shown in Table 5-15. Procedures used to compute all the values given in this table are described in Attachment 7.

(2) Interpretation of Table 5-15. By dividing the forecasts into two probability groups (one equal to or above a selected critical probability and the other below), we have, in effect, created a special type of tailored categorical "yes or no" forecasts. The dividing threshold is the critical probability value rather than the normal 50% probability. The matrices on the left side of Table 5-15 show distributions of the forecasts and occurrences/non-occurrences of the event that could be expected for selected critical probabilities. Also shown in the center and to the right are the overall percent correct, post agreement, and prefigurance for the forecasts in each matrix.

(a) Definition of terms. Tables 5-16a and b translate the matrix values into commonly used operational terms concerned with storm protection or flying. Terms in the first table will be used to explain the critical probability example.

Table 5-14. Distributions of Forecasts and Event Occurrences for Selected Critical Probabilities (Note: Notations "a+b," "c+d," "a," and "c" in Part B pertain to instructions given in Atch 7.)

PART A - FCST VERIFICATION			PART B - CUMULATIVE SUMS			
FCST PROB (P)	TOTAL # OF FCSTS	TOTAL # OF OCCUR-RENCES	# OF FCSTS ≥P (a+b)	# OF FCSTS <P (c+d)	# OF OCCUR-RENCES ≥P (a)	# OF OCCUR-RENCES <P (c)
0	907	19	2208	0	312	0
2	185	9	1301	907	293	19
5	218	15	1116	1092	284	28
10	294	38	898	1310	269	43
20	165	32	604	1604	231	81
30	139	44	439	1769	199	113
40	103	41	300	1908	155	157
50	70	32	197	2011	114	198
60	57	28	127	2081	82	230
70	32	21	70	2138	54	258
80	16	13	38	2170	33	279
90	13	11	22	2186	20	292
100	9	9	9	2199	9	303
ALL	2208	312	2208	2208	312	312

(b) Selecting a Critical Probability. The verification matrices show the customer the real effect a chosen critical probability has on his operation. The largest number of successful launches in Table 5-15 is associated with the lowest critical probability (2%). This P_c also gives the smallest number of missed opportunities; however, those desirable consequences are obtained at the expense of an increase in the number of aborts and a decrease in correct stand downs. In a categorical sense, low critical probabilities result in substantial overforecasting. Normally, only high priority and urgent missions would justify such a low critical probability. Such a low critical probability might be well justified for a severe weather decision problem. At the other extreme, a critical probability of 90% yields the lowest number of expected sorties and largest number of missed opportunities. It also gives the lowest number of aborts and the highest number of correct stand downs. High risk missions (lives, money, political embarrassment, etc) might use a critical probability this large. If the operator stipulates that the number of aborts should not exceed successful launches, a critical probability of approximately 38% (interpolating) would be chosen. Corresponding values of percent correct, post agreement, and prefigurance are included to illustrate the variations in percentages rather than numbers, since some customers may desire this kind of presentation as well.

d. Merits of Using Critical Probabilities.

(1) The obvious advantage of using critical probabilities in decision making is that they are predetermined by the decision maker and appropriate action implemented whenever the critical probability threshold is exceeded. Thus, some follow-on decisions could be made without direct involvement by the decision maker.

(2) Critical probabilities are determined in a number of formal and informal ways. One method, similar to that described, uses simulated forecast distributions. This approach is described in Chapter 6.

(3) The use of monetary value is a good starting point for determining critical probability. However, if actual values are not available, rough approximations are usually adequate. The accuracy of the critical probability need not be any more than one-half the value of the probability intervals used in making the forecasts.

(4) Critical probabilities can be adjusted either objectively or subjectively as priorities and other factors that affect the decision change. For example, a wing commander may establish a critical probability for use when training missions are on schedule, but, if training falls behind schedule, a lower value (depending upon the number of missions needed and time remaining) could be substituted.

Table 5-15. Operational Verification of Selected Critical Probabilities

SEL CRIT PROB	OBSVD	FCST PROB (P)		TOTAL	OVERALL % CORRECT	POST AGREEMENT (% Time Fcst Event Occurred)		PREFIGURANCE (% Time Obsvd Event Correctly Fcst)	
2%		P≥2%	P<2%			P≥2%	P<2%	P≥2%	P<2%
	Yes	293	19	312		22.5	2.1	93.9	6.1
	No	1008	888	1896	53.5	77.5	97.9	53.2	46.8
	Total	1301	907	2208					
5%		≥5%	<5%			≥5%	<5%	≥5%	<5%
	Yes	284	28	312		25.4	2.6	91.0	8.9
	No	832	1064	1896	61.1	74.6	97.4	43.9	56.1
	Total	1116	1092	2208					
10%		≥10%	<10%			≥10%	<10%	≥10%	<10%
	Yes	269	43	312		30.0	3.3	86.2	13.8
	No	629	1267	1896	69.6	70.0	96.7	33.2	66.8
	Total	898	1310	2208					
20%		≥20%	<20%			≥20%	<20%	≥20%	<20%
	Yes	231	81	312		38.2	5.0	74.0	26.0
	No	373	1523	1896	79.4	61.8	95.0	19.7	80.3
	Total	604	1604	2208					
30%		≥30%	<30%			≥30%	<30%	≥30%	<30%
	Yes	199	113	312		45.3	6.4	63.8	36.2
	No	240	1656	1896	84.0	54.7	93.6	12.7	87.3
	Total	439	1769	2208					
40%		≥40%	<40%			≥40%	<40%	≥40%	<40%
	Yes	155	157	312		51.7	8.2	49.7	50.3
	No	145	1751	1896	86.3	48.3	91.8	7.6	92.4
	Total	300	1908	2208					
50%		≥50%	<50%			≥50%	<50%	≥50%	<50%
	Yes	114	198	312		57.9	9.8	36.5	63.5
	No	83	1813	1896	87.3	42.1	90.2	4.4	95.6
	Total	197	2011	2208					
60%		≥60%	<60%			≥60%	<60%	≥60%	<60%
	Yes	82	230	312		64.6	11.1	26.3	73.7
	No	45	1851	1896	87.5	35.4	88.9	2.4	97.6
	Total	127	2081	2208					
70%		≥70%	<70%			≥70%	<70%	≥70%	<70%
	Yes	54	258	312		77.1	12.1	17.3	82.7
	No	16	1880	1896	87.6	22.9	87.9	0.8	99.2
	Total	70	2138	2208					
80%		≥80%	<80%			≥80%	<80%	≥80%	<80%
	Yes	33	279	312		86.8	12.9	10.6	89.4
	No	5	1891	1896	87.1	13.2	87.1	0.3	99.7
	Total	38	2170	2208					
90%		≥90%	<90%			≥90%	<90%	≥90%	<90%
	Yes	20	292	312		90.9	13.4	6.4	93.6
	No	2	1894	1896	86.7	9.1	86.6	0.1	99.9
	Total	22	2186	2208					

Table 5-16a. Operational Terms for Matrix Values Involving Flight Operations.

OBSVD	FORECAST PROBABILITY	
	Favorable	Unfavorable
Yes	Successes, hits, or successful launches.	Lost or missed opportunities.
No	Wasted missions or aborts.	Saved sorties or correct stand downs.

Table 5-16b. Operational Terms for Matrix Values Involving Storm Protection.

OBSVD	FORECAST PROBABILITY	
	Storm	No Storm
Yes	Hits	Unforecast events
No	False alarms	Correct no-storm forecasts

(5) If a customer is opposed to using probability forecasts directly, critical probabilities provide an alternate way of providing tailored categorical forecasts. Rather than using 50% as the threshold for deciding whether or not an event will occur, the critical probability could serve as the threshold. Thus, the resultant decisions will be more cost-effective than conventional categorical forecasts in the long-run.

e. Problems in Using Critical Probabilities.

(1) When the customer's critical probability is outside the limits within which reliable forecasts can be reasonably assured, the customer should be making decisions based on climatology.

(2) Forecasters should not let the value of the critical probabilities influence the value of their forecast probabilities. There may be occasions when a customer changes his critical probability without the forecaster's knowledge.

5-7. Value Analysis.

a. Once a customer's critical probability is determined, yes/no decisions are made based on whether or not the probability forecast exceeds this critical probability. This is a type of categorical forecast based on the critical probability. This is the optimum forecast from the customer's point of view. However, this forecast may not be the most accurate forecast. Table 5-15 is used to illustrate this point. Consider overall percent correct as a measure of accuracy. Note that a categorical forecast based on a critical probability of 70% has the maximum overall percent correct value (87.6%). If the customer's critical probability for this example was 70%, he would have received the most accurate forecast. However, with a critical probability of 30%, the categorical forecasts would not have been as accurate (84% overall percent correct). Therefore, the optimum forecast (based on the critical probability) may not be the most accurate forecast (Kernan, 1975).

b. Effect of Reliability and Sharpness. When the concept of decision models was introduced, one of the assumptions was that the forecasts are reliable; otherwise, errors would occur in the expected costs depending upon the magnitude of the net reliability. It was also stated that, although sharpness is not the main issue in these models, it is important. These effects are illustrated as follows. Consider six sets of forecasts (110 forecasts/set) for a decision problem where rain affects an operation as indicated in Table 5-17.

Table 5-17. Expected Cost Matrix for Rain Protection.

ACTIONS	STATES OF WEATHER		EXPECTED COSTS (E)
	Rain	No Rain	
Protect	\$ 45	\$ 45	$E_1=45P_1+45P_2=\$45$
No Protection	\$100	0	$E_2=100P_1$
PROBABILITY	P_1	P_2	

Since this example fits the cost-loss model, the critical probabiltiy = C/L = 45/100 = .45. The decision rule that would be used is:

Protect if $P_1 > .45$
Indifferent if $P_1 = .45$
No Protection if $P_1 < .45$

Using this model, total costs incurred for each of the six sets of forecasts were calculated (Table 5-18).

Table 5-18. Effect of Sharpness and Reliability on Expected Costs

PROBA-BILITY (P)	# OF FCSTS(n)	PERFECTLY RELIABLE			MODERATELY RELIABLE		
		# OCCUR-RENCES(D)	OBSVD FREQ %	ACTUAL COSTS(E)	#OCCUR-RENCES(D)	OBSVD FREQ	ACTUAL COSTS(E)
LITTLE SHARPNESS							
100	10	10	100	$ 450	9	90	$ 450
90	10	9	90	450	9	90	450
80	10	8	80	450	9	90	450
70	10	7	70	450	7	70	450
60	10	6	60	450	5	50	450
50	10	5	50	450	4	40	450
40	10	4	40	400	5	50	500
30	10	3	30	300	5	50	500
20	10	2	20	200	2	20	200
10	10	1	10	100	0	0	0
0	10	0	0	0	0	0	0
ALL	110	55	50	$3700	55	50	$3900
MODERATE SHARPNESS							
100	30	30	100	$1350	25	83.3	$1350
50	30	15	50	1350	20	66.7	1350
40	15	6	40	600	8	53.3	800
20	20	4	20	400	2	10.0	200
0	15	0	0	0	0	.0	0
ALL	110	55	50	$3700	55	50.0	$3700
PERFECT SHARPNESS							
100	55	55	100	$2475	50	90.9	$2475
0	55	0	0	0	5	9.1	500
ALL	110	55	50	$2475	55	50.0	$2975

(1) These examples assume that a decision was made (protect or no protection) for every forecast (110) in each set using the decision rule above. Costs are totals for each probability interval. For example, if there were 10 forecasts in the interval and the decision cost associated with each forecast was \$45, then the total cost was \$450.

(2) Three sharpness patterns are shown to illustrate the effect this attribute has on actual, versus expected costs. For each sharpness example there is a set of forecasts with perfect reliability and another moderately reliable.

(3) In every case, the actual costs equal the expected costs, whenever the forecast probability (P) is greater than the cost-loss ratio (C/L = 45%). This is true regardless of how reliable the forecasts are. The reason is that the protection costs are fixed at \$45 per decision (forecast), and that cost is unchanged, whether the event occurs or not.This is seen by inspecting the costs in all six examples where P > 45%.

(4) We can examine the variation of actual costs due solely to sharpness by considering only the data for perfectly reliable forecasts in Table 5-18. The total actual cost was \$3700 for perfectly reliable forecasts with little sharpness, compared to a total actual cost of \$2475 for perfectly reliable forecasts with perfect sharpness. A similar comparison can be made for moderately reliable forecasts. If we compare the total actual costs for perfectly and moderately reliable forecast with the same degree of sharpness we see the variation of actual costs due solely to reliability (\$3700 vs \$3900 for little sharpness). Overall, the total cost for the moderately reliable forecasts is \$200 more than the actual/expected cost, \$3700, for the perfectly reliable set. The total number of occurrences of the event is the same in both sets of forecasts, but the difference in costs exists because the unreliable forecasts have two additional event occurrences in the intervals below the critical probability than indicated by perfect reliability.

(5) The moderately sharp, perfectly reliable set of forecasts, when compared with the moderately sharp, moderately reliable set, show no change in the actual/expected cost from the perfectly reliable, little sharpness set of forecasts. In this case, underforecasting in the 40% interval is offset by the overforecasting in the 20% interval. The reason the three sets of forecasts cost the same is that the distributions of the number of forecasts and event occurrences above and below the cost-loss ratio are identical. Thus, differences in sharpness or reliability have no effect on decision costs, unless they redistribute the forecasts and occurrences across the cost-loss ratio.

(6) The last two sets of figures illustrate the above point. The group on the left are perfect forecasts, and represent the lowest possible cost the decision maker could expect in conducting the operation. The costs are lower, because the customer took protective action only on those days when the event occurred, and the damage loss was zero. Although the set on the right was perfectly sharp, reliability errors increased the cost due to damage. The cost, however, was still lower than all others, except for perfect forecasts.

(7) Summary. The conclusion from these examples is that both sharpness and reliability affect the decision costs. The extent to which they affect costs depends upon the distribution of forecast and event occurrences with respect to the critical probability (cost-loss ratio).

c. Value of Weather Forecasts. The probability forecasts and the models described enable one to calculate the relative values of forecast information.

(1) Climatological forecasts. In the absence of a forecast, the decision maker can always use the climatological probability to determine the best course of action. Using the expected cost matrix given earlier in Table 5-17 and the sample climatology (C_j - 50%) from Table 5-18 (55 occurrences/110 forecasts) the expected cost (E(CLIM)) per decision for that operation is calculated as follows:

$$E_1\,(CLIM) = C \cdot \$45$$

$$E_2\,(CLIM) \cdot L\,P_1 = \$100\,(.5) \cdot \$50 \qquad (5\text{-}13)$$

where $P_1 = C_j$

Thus, by using only the climatological probability, the customer would be better off to take protective action each time a decision is made, since it results in the least cost over the long-term. For other cases, calculation of E1 (CLIM) and E2 (CLIM) can be accomplished by using the generalized matrix described earlier in Table 5-10. The total cost, E_T(CLIM) for the set of forecasts in Table 5-18 which has little sharpness and perfect reliability is calculated by multiplying the unit costs by the total number of forecasts (N) where N = 110. The general equation follows.

$$E_T(CLIM) = \begin{cases} N(E_1(CLIM)) & \text{if } C_j \geq P_c \\ N(E_2(CLIM)) & \text{if } C_j < P_c \end{cases} \qquad (5\text{-}14)$$

Where C_j = sample climatological probability

P_c = critical probability or cost-loss ratio, C/L

N = total number of forecasts in the set

Since $P_c = C/L = .45$ and $C_j = .5$, $C_j \geq P_c$, the total cost is:

$$E_T(CLIM) = N(E_1\,(CLIM))$$

$$E_T\,(CLIM) = \$110\,(45) = \$4950 \qquad (5\text{-}15)$$

This is a significant savings over the cost that would have occurred had protective action not been taken. By using the other rule in equation 5-14, that cost would have been \$5500.

(2) Probability forecasts. Similar total costs can be calculated for probability forecasts as illustrated earlier in Table 5-18. For the set of forecasts which has little sharpness and perfect reliability, the total cost is E_T (PROB) = \$3700.

(3) Categorical forecasts. The unit and total costs can also be calculated based on categorical forecasts. Procedures are identical to those used for calculating costs for probability forecasts, with one exception. Instead of using the critical probability to transform the probability forecast into categorical forecasts, 50% or some other realistic probability value is

used. Assuming a threshold of 50%, the total cost for the set of forecasts in Table 5-18 (little sharpness and perfectly reliable) is E_T (CAT) = \$3750.

(4) Perfect forecasts. Calculation of total costs for perfect forecasts is also illustrated in Table 5-18 (set which was perfectly sharp and perfectly reliable). That cost is E_T (PERF) = \$2475.

(5) Value Comparisons. The true value of forecasts to the customer is found by comparing expected costs associated with different forecasts to the cost that would have been incurred if only climatology had been used to make the decision. This latter cost represents the upper bound of the cost. The value of each set of forecasts is the difference between the cost from using that set and the cost from using climatology. The lower bound is given by the cost from using perfect forecasts.

$$V(PROB) = E_T(CLIM) - E_T(PROB) = 4950 - 3700 = \$1250$$
$$V(CAT) = E_T(CLIM) - E_T(CAT) = 4950 - 3750 = \$1200$$
$$V(PERF) = E_T(CLIM) - E_T(PERF) = 4950 - 2475 = \$2475 \qquad (5\text{-}15)$$

In many cases, monetary values will not be available for computing expected costs. If utility values exist, however, they can be used to indicate the expected values.

(6) Summary. Murphy (1976e) performed an empirical study of the relative value of climatology, categorical, probabilistic, and perfect forecasts in the cost-loss situation. He concluded that the expense (value) associated with perfectly reliable probabilistic forecasts *is* less (greater) than or equal to the expense (value) associated with climatological and categorical forecasts for all values of the cost-loss ratio, C/L.[1] For unreliable probability forecasts, the expense (value) may be greater (less) than the expense (value) associated with climatological and/or categorical forecasts for some values of C/L. However, an examination of a number of samples of unreliable probability forecasts indicates that the first relationship (for reliable forecasts) appears to hold true for most (if not all) values of C/L, even for moderately unreliable forecasts. Moreover, the study suggests that if the value of unreliable probability forecasts is exceeded, it will be by the value of categorical forecasts. Murphy finally concludes that the value of the meteorological product can be significantly increased if probability forecasts for a variety of weather conditions are routinely formulated and disseminated to decision makers, including the general public.

5-8. Other Models. The preceding discussions presented the most common and simple decision models that have been successfully applied to meteorological decision problems in the past. There are others, but they are too difficult to be of value in this pamphlet. Brief descriptions of other models follow, so you will know of their existence and can avoid tackling unique decision problems without the proper tools.

a. Tactical and Strategic Decision Model. Many types of problems occur where decisions are made sequentially on the basis of a continuing flow of weather information. A two-stage decision model could be developed in which the decision maker must first make a strategic decision regarding the amount of protection that can be obtained and kept available for subsequent use. This decision would be followed by a tactical decision whether or not to employ a certain amount of protection on a particular occasion (Murphy, 1976a). This model could be applied where there are several degrees or types of protective action. Consider the base civil engineers (CE) and their snow removal plan. Not only is the probability of snow important, but so is its amount and intensity. In a situation where there is a low probability of a light to moderate accumulation, CE may only check their equipment, sand, and salt supplies and place a small force of workers on home alert. For a higher probability of moderate accumulation, key supervisors and a small work force may be recalled; other workers may be placed on alert, equipment and supplies may be positioned, and actual clearing started only if snow has begun. On the other hand, a high probability of heavy snow might mean total recall and immediate commencement of clearing action (Nelson and Winter, 1960).

b. Two-Way Call Model. This model is a variation of the basic two-stage model in which there are two separate courses of action available. The variation is actually a hedging operation consisting of adding a third course of action, which is simply a delay until the last minute in deciding between the first two actions. Delaying the decision (action 3) adds cost, but when the probability forecast is near the critical probability, the third course of action, in some cases, is more cost effective in the long run. This is especially true when some of the resources can be used in *either* of the first two actions. This particular model has possible applications in launch decisions, severe weather protection, etc, whenever the customer desires that provisions for last minute decisions be built into the model (Nelson and Winter, 1960). Note: The simpler two-stage model can still be used with these same decisions, when the built-in delay option is not required.

c. Linear Postponement Model. This model involves decisions where there are two choices: to attempt a job, or delay and accept a penalty. This model is best described by assuming that the cost of completing a job can be broken down into the following three elements: the direct cost of doing the work, a fixed cost or penalty charge for each day that elapses before the job is complete, and an added loss incurred each day the job is started, but unfavorable weather results (Nelson and Winter, 1960). Construction decisions readily fit this model, but it could also be applied to training schedules and other types of decisions.

d. Postponement Model. This model is a variation of the linear postponement model. Instead of having an indefinite period of time in which to complete the job, the decision maker must finish it by a given deadline, or else incur a penalty. The penalty might be a full or partial refund of any gross revenue paid the decision maker, who is no longer required to complete the job. This case would arise if completion of the job after the deadline provided no value to the agency letting the contract. The

[1] Actually, E(PROB) = E(CLIM) only when C/L = 0 and 1; and E(PROB) = E(CAT) only when C/L equals the sample climatology.

penalty might also be the expected cost of eventually completing the job, with no deadline, but with some higher cost or penalty applying after the deadline. A variety of other penalty combinations might also be used (Nelson and Winter, 1960). Applications of this model are similar to the linear postponement model.

e. Summary. Models attempt to develop objective rules which reproduce the decision maker's thought process. Consequently, the model chosen must be matched to both the decision maker and the decision problem. In addition to the examples cited, there are specific models designed for decision makers who are inclined to take risks or for those who wish to avoid risks. Unfortunately, there is little information published on application of decision models to the many military weather decision problems our customers face. Thus, there will undoubtedly be situations where we will have to develop or adapt models to handle unique decision problems.

Chapter 6

INTRODUCTION TO WEATHER IMPACT AND MISSION SUCCESS INDICATORS

6-1. Introduction. In this chapter, we introduce the concept of a Weather Impact Indicator (WII) and how it is used to calculate a Mission Success Indicator (MSI). The main forms of WIIs and how to construct them are discussed. The different weather effect models used by the customer are described, with the form of WII which can be used to support each one as input to an MSI. The calculation of MSIs is covered in some detail, including some discussion of non-weather effects. While the customer will calculate his MSI, it is essential that the staff weather officer have an intimate knowledge of the mission and problems involved. This understanding can be used to better tailor the support provided. A WII may be produced by methods other than a forecast. These different methods are discussed in the final section.

6-2. Forms of WIIs. A WII is the probabilistic weather input used to calculate an MSI. WIIs are tailored for specific decisions. They can be calculated for an overall mission, or for any particular stage of a mission, e.g., take-off, enroute, aerial refueling, weapons delivery, and recovery.[1] There are two main forms of WII, a threshold forecast and a continuous probability distribution.

a. **Threshold Forecast.** This is the simplest of the two forms. It is the probability that the weather will exceed a particular threshold value (ceiling above 1500 feet, winds greater than 5 knots, temperature below freezing, etc.), or that an event will or will not occur (rain, thunderstorms, freezing rain, hail, etc.).

A categorical forecast is a special form of this type forecast; one in which only probabilities of 100% or zero are inferred. This is typical of our normal weather support, but does not convey all of the information possible. For example, consider some operation to take place at 0930L where wind in excess of 17 knots is a critical factor (See Table 6-1.)

Note the categorical forecast, as might be given on a terminal forecast. This says that wind gusts will exceed 17 knots during a time period covering the operation. We still don't know how often, especially near a particular time, or how sure the forecaster is.

The general probability forecast represents the all-purpose, area forecast available from a weather central. The nature of such a generalized forecast - valid over an area, an interval of time, and a different threshold - degrades it's application to the specific operation.

A probability forecast tailored to the specific threshold (17 knots), time (0930L), and location best meets the customer's requirements for decision making. This is the threshold forecast form of WII.

Subjective threshold forecasts are relatively easy to make. Given a weather element/threshold, the forecaster examines the relevant observations, analyses, forecasts, and climatology. Based on past experience, the forecaster subjectively estimates the probability of the weather exceeding the threshold. This process is basic to every manual forecast, whether expressed in categorical or probabilistic terms.

b. **Continuous Probability Distribution.** Suppose a forecaster gives a 60% probability of winds exceeding 15 knots for a particular weather situation. This same forecaster is then asked for the probability of winds greater than 20 knots for the same situation. Will his probability forecast for this threshold be higher, lower, or the same? What if the threshold is 25 knots, 50 knots, or 100 knots? The probability for exceeding a higher threshold is less than that for any lower threshold. Eventually, the probability for exceeding a specific high wind speed becomes zero. Since the probability for the wind speed being zero or greater is 100%, we see that there is a continuous distribution of forecast probability versus the wind speed threshold. We may present these distributions in two ways, as a cumulative probability curve or as a probability density curve.

(1) **Cumulative probability curve.** An example of a cumulative probability curve is shown in Figure 6-1. Each point on the curve gives the probability for wind speed at or less than the value on the horizontal axis. The probability for a wind speed of 15 knots or less is shown by the dashed line as 40%. Cumulative probabilities for other thresholds are found from the curve in a similar manner. These curves may be obtained through subjective or objective forecasts.

(a) **Subjective forecasts.** Cumulative probability curves can be generated subjectively for any continuous weather element - ceiling, visibility, wind speed, temperature, etc. - for a location and forecast time. A forecast for a single threshold represents one point on the curve. Probability forecasts for a series of thresholds of an element can be plotted on a graph similar to figure 6-1 and the points connected to form a "complete" cumulative probability curve. Figure 6-1 was constructed in this manner by using the threshold forecasts: 0 kts - 0%; ≤ 5 kts - 5%; ≤ 10 kts - 15%; ≤ 15 kts - 40%; ≤ 20 kts - 70%; ≤ 25 kts - 85%; and ≤ 30 kts - 95%.

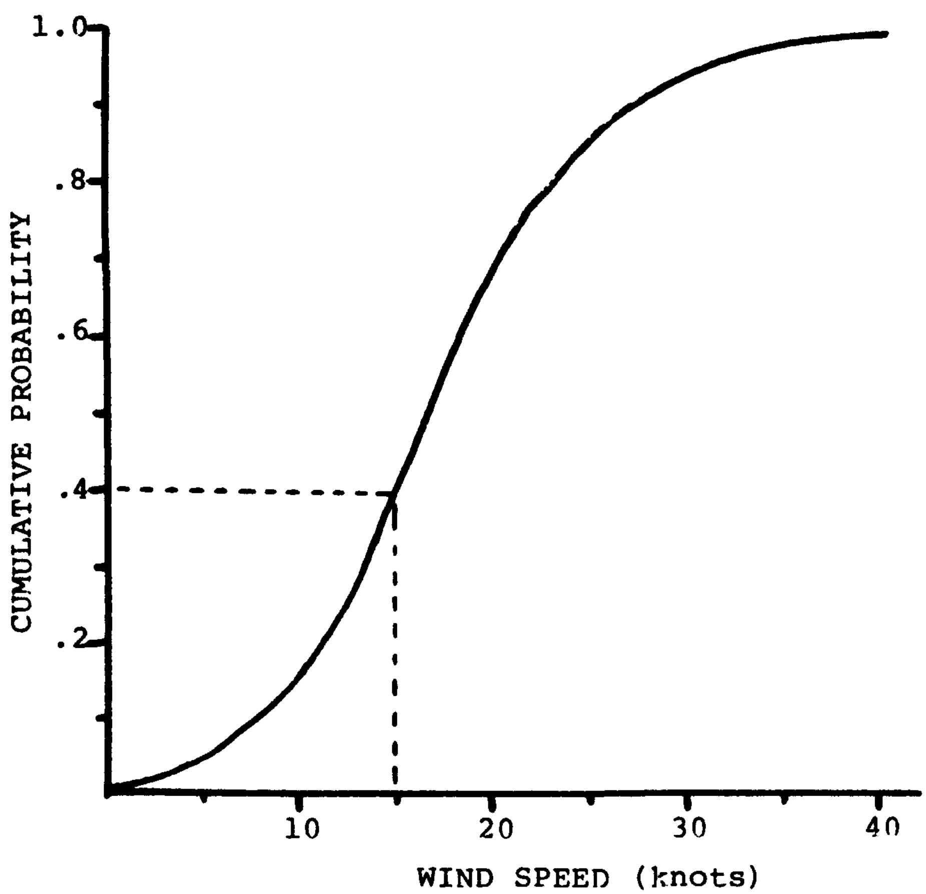
Several experiments have been conducted where forecasters predict the cumulative probabilities for the maximum and minimum temperature (Peterson, Snapper, and Murphy, 1972; Murphy and Winkler, 1974b; Murphy and Winkler, 1975; Murphy and Winkler, 1977). In one approach, forecasts were made by successive division of the temperature range into equal probability ranges. A detailed discussion of this procedure is given in Attachment 10. Another approach tasked forecasters to assign probabilities that the temperature maximum (minimum) would be within a fixed temperature interval (5 or 9°F). These experiments have shown that experienced forecasters can reliably describe the uncertainty inherent in their temperature forecasts. Further forecast experiments at the Massachusetts Institute of Technology (Sanders, 1973; Sanders, 1976) have shown that inexperienced students can produce reasonable probability forecasts for minimum temperature in ten intervals about the climatic mean and six categories of precipitation amounts for four consecutive 24-hour periods.

[1]Special techniques are required to compute the combined WII to account for the spatial and temporal correlations of weather.

Table 6-1. Forecast Examples.

CATEGORICAL FORECAST	GENERAL PROBABILITY	WII
00-12L Wind 10G20	00-12L Gust >15 80%	0930L Gust >17 50%

Figure 6-1. Cumulative probability of wind speed.



These subjective techniques can be applied in predicting the probability distribution for any continuous meteorological variable - visibility, ceiling, wind speed, etc. The main requirements, other than basic knowledge, are practice and feedback of verification results.

Uncertainty in a forecast increases with time. The effect of this on a cumulative probability curve is shown in Figure 6-2. Curve A represents the distribution for a short range forecast. The curve indicates a high certainty for a wind speed near 17 knots. The cumulative probability of curve A increases from about 20% at 15 knots to about 85% at 20 knots. Thus, the probability of winds between 15 and 20 knots is about 65%. (85% - 20% = 65%) Curve B represents a medium range forecast. The value of B increases more gradually than the value of A. This indicates that the probability distribution is broader, and the forecast less certain for any given interval of speed. Curve C might be the cumulative probability distribution for a long range forecast or for climatology.

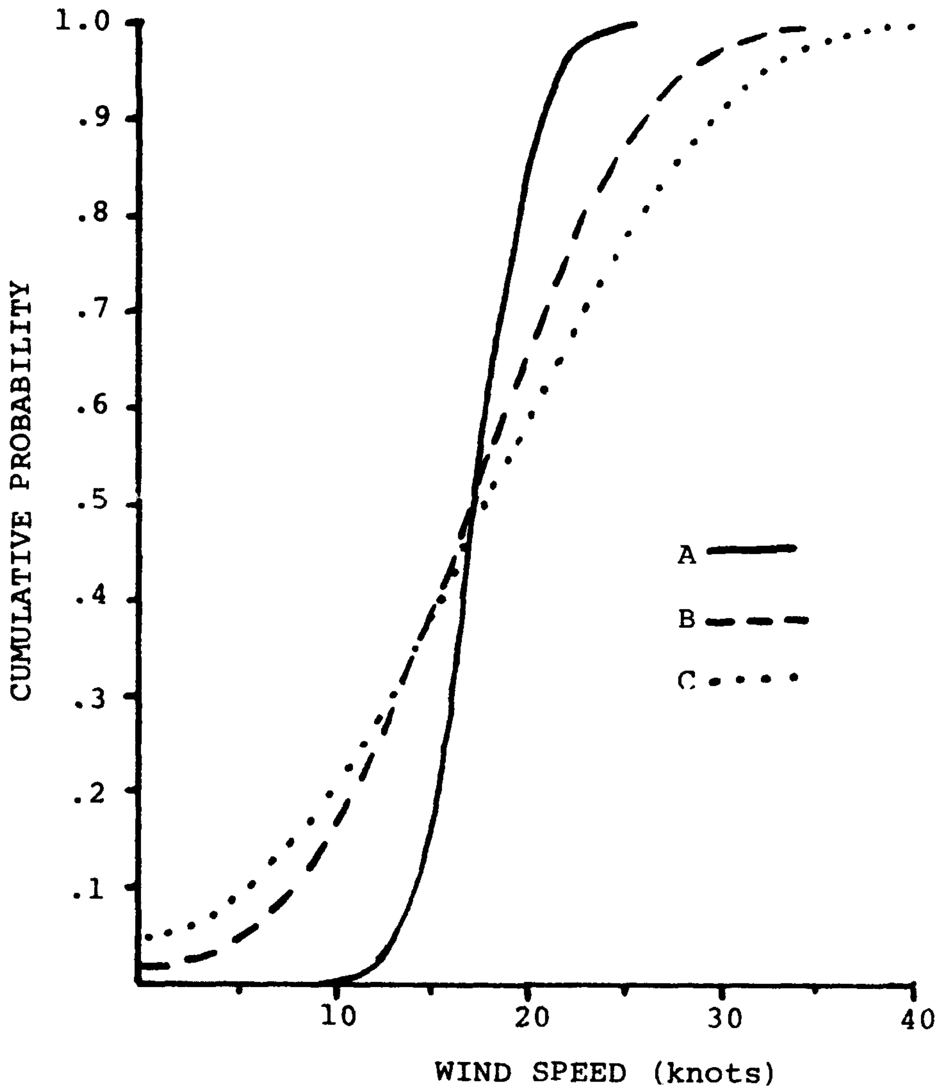


Figure 6-2. Effect of time on uncertainty of the forecast.

As uncertainty increases in a subjective forecast, the distribution should approach the one for climatology as a "no skill" base. Climatological cumulative probability curves for many elements can be derived from data in a RUSSWO. Table 6-2 gives the wind speed frequency for all wind directions and weather categories from the Ft Rucker RUSSWO for March, 1200-1400 LST. The cumulative frequencies are also given, and a plot of these is shown in Figure 6-3. The probability of winds less than or equal to 15 knots from Figure 6-3 is about 93%, thus giving a climatological probability of 7% for mean winds greater than 15 knots. Similar plots can be made using RUSSWO information for ceiling, visibility, temperature, precipitation, and sky cover for use as a base in constructing subjective forecasts.

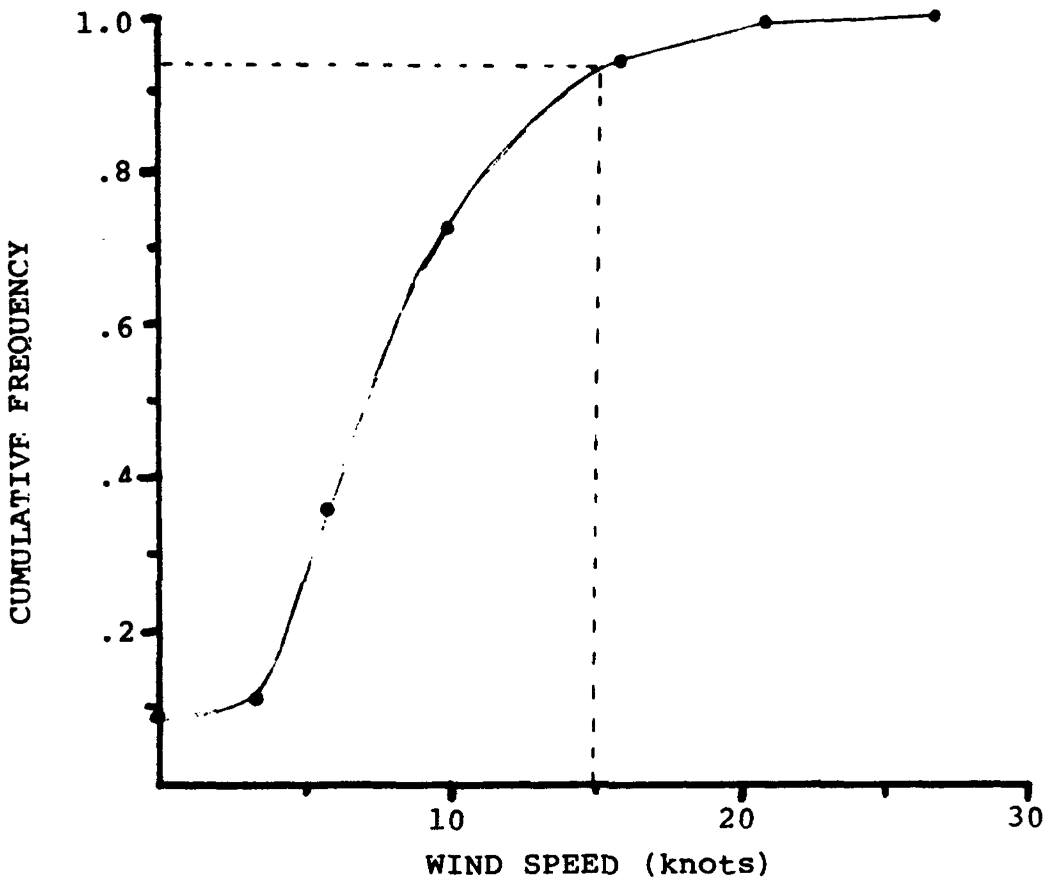


Figure 6-3. Cumulative frequency plot for the data in Table 6-2.

Table 6-2

PERCENT FREQ WIND SPEED

Calm	1-3 Kts	4-6 Kts	7-10 Kts	11-16 Kts	17-21 Kts	22-27 Kts
4.7	6.5	24.8	36.8	22.0	4.8	0.4

CUMULATIVE FREQ WIND SPEED

4.7	11.2	36.0	72.8	94.8	99.6	100.0

(b) **Objective forecasts.** Objective methods may be used to construct the cumulative probability distribution. One of the easiest ways is to use data from the MOS bulletins. (See NWS Tech Procedures Bulletin #217, 3 Nov 77, for category definitions and bulletin format.) The MOS bulletin probabilities are given to the nearest 10%, and may sum over all categories to more or less than 100% at times due to rounding. Some judgement must be exercised when plotting the cumulative sums due to this rounding. Figure 6-4 illustrates how three forecast ceiling probability distributions from a MOS bulletin are plotted in cumulative distribution form from category boundaries of 200, 500, 1000, 3000, and 7500 feet.

Note how the forecast distributions in Figure 6-4 become less sharp as the length of the forecast increases and that, with time, they trend toward climatology, i.e., a high ceiling or no ceiling.

Other methods may use a subjective input for some value which then determines an entire distribution by objective means. One method developed by AWS/DN is the Multi-Category Probability Variation Guide algorithm. This algorithm produces the probability distribution for all thresholds of a weather element, given the length of the forecast, the climatological distribution of the forecast element, and a forecast probability for exceeding one threshold. Since this method can prove quite useful, the use of it will be covered briefly, with some examples of the output.

The examples shown are based on a six hour forecast for Scott AFB, valid at 1800Z, in December. The Scott AFB climatology for that time for the four AWS ceiling categories and the length of the forecast are required inputs to the program. Two methods of obtaining forecasts are possible.

The first way to use the algorithm is to subjectively predict the probability of exceeding a single threshold. In Table 6-3a, the prediction was made for exceeding category C, i.e., being in category D. For each value entered in the column under D, the algorithm produced all the entries for the other categories. If 50% probability is forecast in category D, then the probability for C is 44% and for B is 5%. Note that round-off causes the sums to occasionally differ from 100%.

The second way to use the algorithm is to rank the weather situation as to the degree that it favors "good" weather. The ability to do this reliably is a function of forecaster experience. Forecasters quickly learn to recognize "bad" and "good" weather situations from routine forecasting aids; in effect, developing a mental file of map types associated with the expected weather conditions for their area. When evaluating the current situation, the forecaster mentally compares it with past experience and, with a little thought and practice, ranks it on a scale of 1 to 99, worst case to best case. This rank, expressed as a percentage, is used as an input to the algorithm.

Table 6-3b shows examples of the output for different rank inputs. If the forecaster believed that the situation was an average one with a rank of 50%, a 15% probability for category C and an 85% probability for category D would be read. The results should not be surprising in view of the ceiling climatology and the shortness of the forecast period (high skill). When longer forecast periods are used, thus assuming lower skill, the forecasts would converge toward climatology.

Tables 6-3a and 6-3b are only two examples of the use of the algorithm. We could use other elements; a forecast for one time period can be used to also provide the forecast for another time period; or an option is available to "spread" a probability forecast to find the forecast at nearby locations.

Tables based on this algorithm are available from USAFETAC for use as forecast aids. The algorithm can be run on a handheld calculator, even a rather small one. The program can be obtained through AWS/DN.

(2) **Probability density curve.** The second method of presenting a continuous probability distribution is through a probability density curve. This curve is directly related to the cumulative probability curve; the probability density is the slope (derivative) of the cumulative probability curve. An example is shown in Figure 6-5. This curve is the plot of the slope, i.e., d(cumulative probability)/d(wind speed), of the curve in Figure 6-1, plotted as a function of wind speed.

The total area under the curve in Figure 6-5 is 1.0, or 100% probability. The maximum of the curve is at about 15 knots. This is the "most probable" wind speed. The probability that the speed will be less than or equal to 15 knots is shown by the shaded area under the curve to the left of the dashed vertical line at 15 knots. This is the integral of the curve from 0 to 15 knots, i.e., $p = \int B'ds$, where s is wind speed and B' is the probability density of wind speed. In this case it is 0.4 of the total area, or 40%. Note that the threshold for 50% probabilty does not necessarily coincide with the most probable wind speed. As the threshold is increased, the area under the curve to the right of the threshold decreases; thus, the probability of exceeding the threshold decreases.

The degree of skill, or certainty, in a probability density distribution is shown by the height of the peak and the spread of the distribution. This is illustrated in Figure 6-6. The three curves, A', B', and C', correspond to A, B, and C in Figure 6-2. They are the derivatives of A, B, and C, respectively. The total area under each of the curves in Figure 6-6 is equal to 1.0. The most probable wind speed for each curve is the same, 17 knots, but the distributions are greatly different. This reflects the uncertainty encountered as the length of the forecast period increases.

The form of WII used will depend on the particular need, or weather effect model, of the customer. Cumulative probability curves are perhaps a more natural way of expressing the probability distribution of a weather element. Certainly they are more easily determined by subjective methods. Simple threshold forecasts are inherent in that distribution. Probability density curves can be derived by graphically differentiating the cumulative probability curve. Whatever technique is used to formulate the forecast probability distribution, such a distribution gives the maximum amount of information about the expected weather.

6-3. Weather Effect Models. The customer can describe the effect of weather on an operation in one of two ways - a simple threshold model where a particular value of a weather parameter forms the decision point, or a continuous function model where the effect of weather varies with the value of the weather parameter.

a. **Simple threshold model.** Simple thresholds are part of everyday weather support, and are

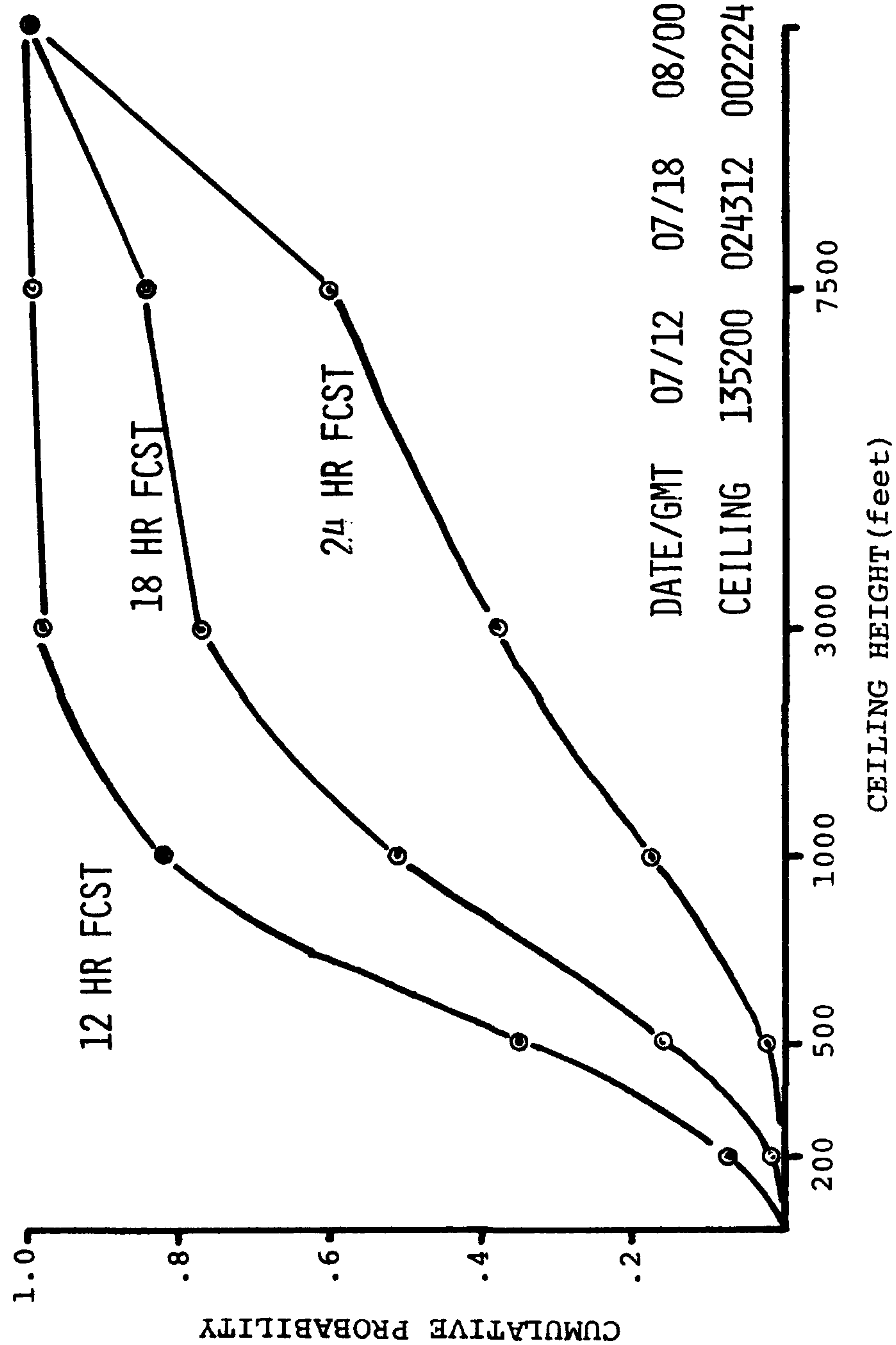


Figure 6-4. Sample cumulative probability forecasts using MOS.

Table 6-3a

Six Hour Ceiling Forecast for Scott AFB, Valid 1800Z
Multi-Category Probability Variation Guide

	A	B	C	D
	52	48	0	0
	1	75	23	1
	0	66	31	2
	0	51	44	5
	0	36	54	10
	0	21	59	20
	0	13	57	30
	0	8	52	40
	0	5	44	50
	0	3	37	60
	0	2	28	70
	0	1	19	80
	0	0	10	90
	0	0	5	95
	0	0	1	99
	0	0	0	100
Climatology:	0.3	10.5	20.7	68.5

Table 6-3b

Six Hour Ceiling Forecast for Scott AFB, Valid 1800Z
Multi-Category Probability Variation Guide

RANK	A	B	C	D
1	7	89	4	0
2	3	87	10	0
5	0	68	30	2
10	0	41	51	8
20	0	14	58	28
30	0	5	44	51
40	0	1	28	71
50	0	0	15	85
60	0	0	6	94
70	0	0	2	98
80	0	0	0	100
90	0	0	0	100
95	0	0	0	100
99	0	0	0	100
Climatology:	0.3	10.5	20.7	68.5

Rank is the degree, in percent, that the situation favors higher categories.

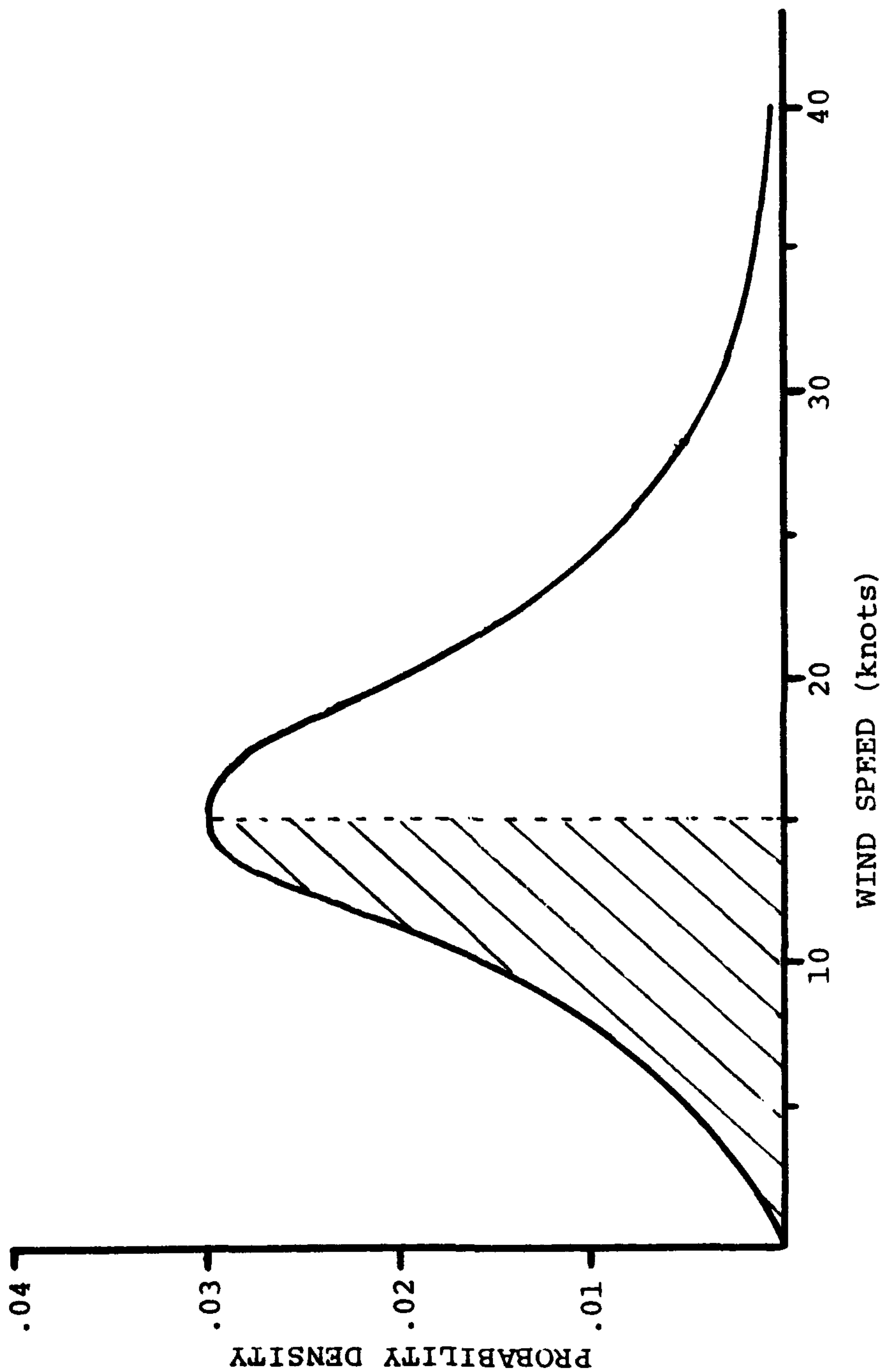


Figure 6-5. Example of a probability density curve.

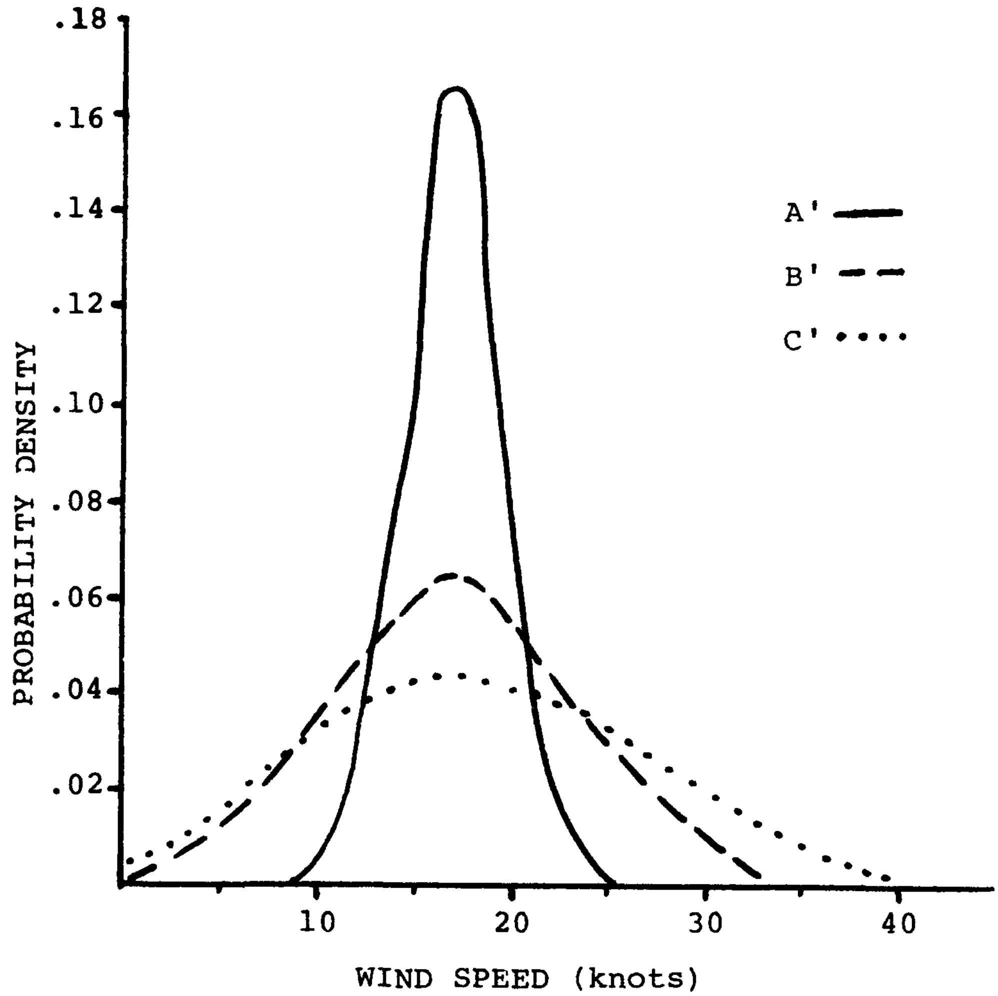


Figure 6-6. Probability density curves for different length forecasts.

commonly used to define "go/no-go" decisions. For example, VFR, PAR, VOR, and TACAN landing minima are discrete ceiling and visibility values; paradrop of personnel is conducted only if the winds are less than 13 knots; and most aircraft have restrictions based on fixed crosswind speeds and/or gust spreads that "prevent" takeoff and landing.

Weather threshold values are often used to ensure a high level of safety for a particular operation. Pilots are given the level of training that enables them to safely land their aircraft 99+% of the time when the ceiling/visibility is greater than or equal to 200/1/2. Highly experienced, skilled pilots, current in the operation of their aircraft, may be able to land successfully 75% of the time, with ceiling/visibility as low as 100/1/8. However, from a safety and economics standpoint, the 25% failure rate is unacceptable and the 200/1/2 threshold is established below which landings are not made.

A simple threshold model is illustrated in Figure 6-7. Note that the scale on the vertical axis has been omitted. System effectiveness can be less than 100%, even in perfect weather.

The WII needed by the customer to use this weather effect model is a simple threshold forecast for the critical weather event. The cumulative probability curve could be used, since the particular threshold forecast needed is just a point on the curve. This has the added advantage of being able to support a customer with multiple thresholds or several customers with different thresholds with one forecast of the weather event.

b. **Continuous function model.** Simple thresholds are not usually realistic descriptions of a system's weather sensitivities. Basing decisions on weather being above or below a single threshold is more a matter of establishing an identifiable limit for conducting operations, rather than any sudden degradation of the system capability when the threshold is just exceeded.

A more general case is one where all goes well when the weather is above one threshold, and complete mission failure results when the weather is below another threshold. Between the two thresholds, the probability of mission success changes as a continuous function of weather. This variation is illustrated in Figure 6-8. Again, the absence of values on the vertical scale is deliberate, since the system may have a probability of success in perfect weather of less than 100%.

A simple threshold forecast form of WII cannot be used to support this model. A continuous probability distribution form of the WII must be used to describe the weather forecast across the range where weather is a factor in the success of the operation.

c **Establishing weather effect models.** Realistic continuous function (multiple threshold) models are far more representative of systems capabilities than simple threshold models. They are also far more difficult to establish. Staffmets and SWOs must work with their customers to determine the model that best reflects system capabilities over all ranges of weather conditions. Careful analysis of weapons delivery results at tactical ranges, the fraction of cloud cover on reconnaissance photos, successful refueling hook-ups, paradrop injury rates, etc., as a function of weather will help in establishing the weather effects for a system. The customer's operations analysis, evaluation, and planning staffs are good places to start in conducting such analyses.

Several iterations may be required before the customer validates the model for the system and/or tactics. The models must be developed before the system is employed in a conflict, if such employment is to be optimal. Once developed, these models will also result in increased effectiveness of weather support to routine training and peacetime operations.

6-4. Mission Success Indicators. The WII furnished to the customer is used, together with the customer's weather effect model, to calculate the impact of weather on a mission. This then forms a part of the Mission Success Indicator (MSI). An MSI is the probability that a mission will succeed. MSIs may be calculated for an entire mission, or for any stage of a mission where a decision option exists. It incorporates the impact of all factors that affect mission accomplishment. These include weather elements, such as ceiling, visibility, crosswind, etc., and non-weather considerations, such as maintenance status, enemy defenses, weapon system kill efficiency, tactics, target type, etc.

Several examples will be used in this section to illustrate the use of a WII to calculate an MSI. The examples will be presented considering only weather effects, then some discussion will be given on how non-weather factors enter into the decisions.

a. **Weather effect only.**

(1) **Equipment paradrop example.** A critical piece of equipment, a radio for command and control, is needed at a forward area. The wing commander plans to paradrop the radio from a C-130 at 0930L. If the surface wind exceeds 17 knots, there is a 10% chance that the radio will be damaged. Using the forecast WII in Table 6-1, what is the MSI for this simple threshold model?

The probability of damage to the radio is the conditional probablity of damage given winds in excess of 17 knots (10%) times the probability of those winds (50%). Thus, the probability of damage is 0.1 x 0.5 = 0.05, or 5%. The probability of success is the probability that the radio will be undamaged, or 1 - 0.05 = 0.95, or 95%.

Suppose that the following information was available from a series of experiments on the effectiveness of equipment packaging for paradrops:

Wind	Damage
Speed ≤15 knots	No damage.
15 < Speed ≤20 knots	10% chance.
20 < Speed ≤25 knots	40% chance.
25 < Speed ≤30 knots	70% chance.
Speed >30 knots	90% chance.

This information is portrayed graphically in Figure 6-9.

We now have a distribution of probability of damage as a function of wind speed. A simple wind speed threshold forecast is obviously inadequate here. We need a forecast covering the speed regime where we are given a probability of damage. The cumulative probability of wind speed given in Figure 6-1 is the WII for this case.

Since the probability of damage is given in discrete

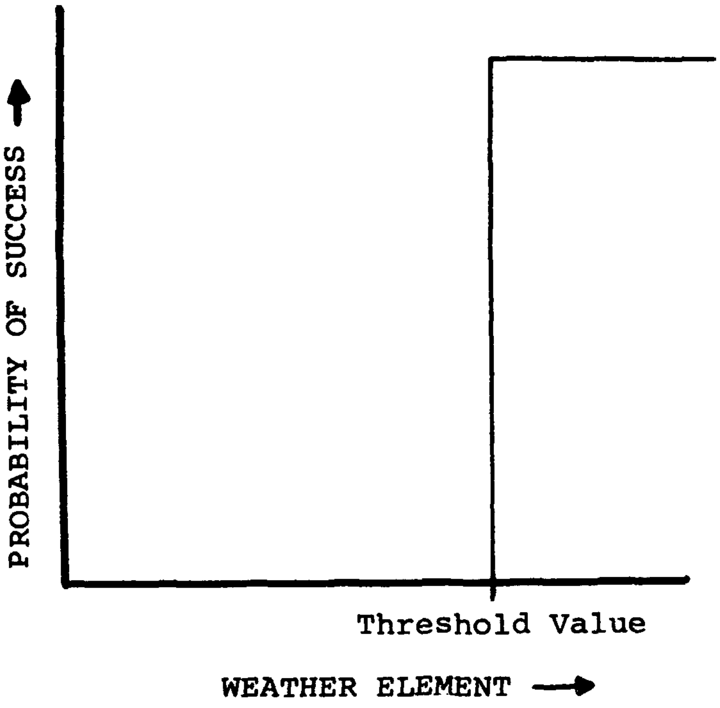


Figure 6-7. Simple threshold weather effect model.

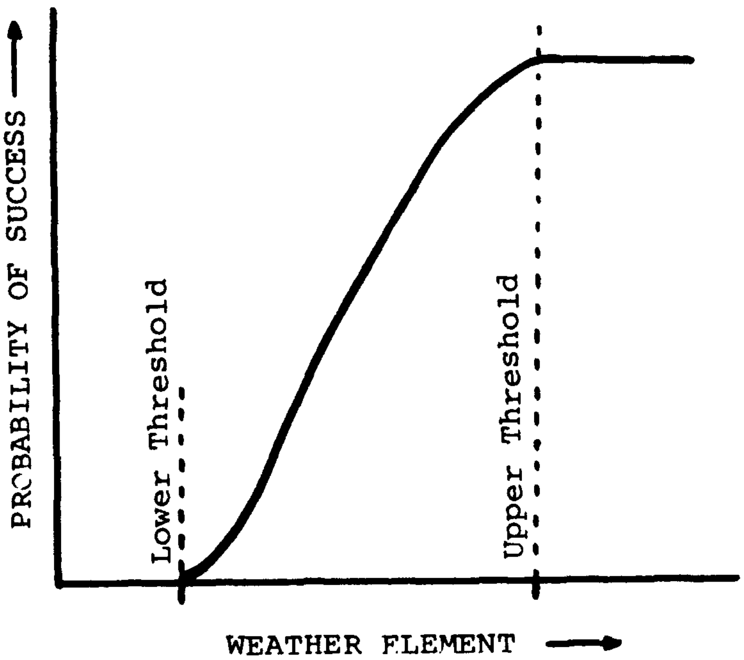


Figure 6-8. Continuous function weather effect model.

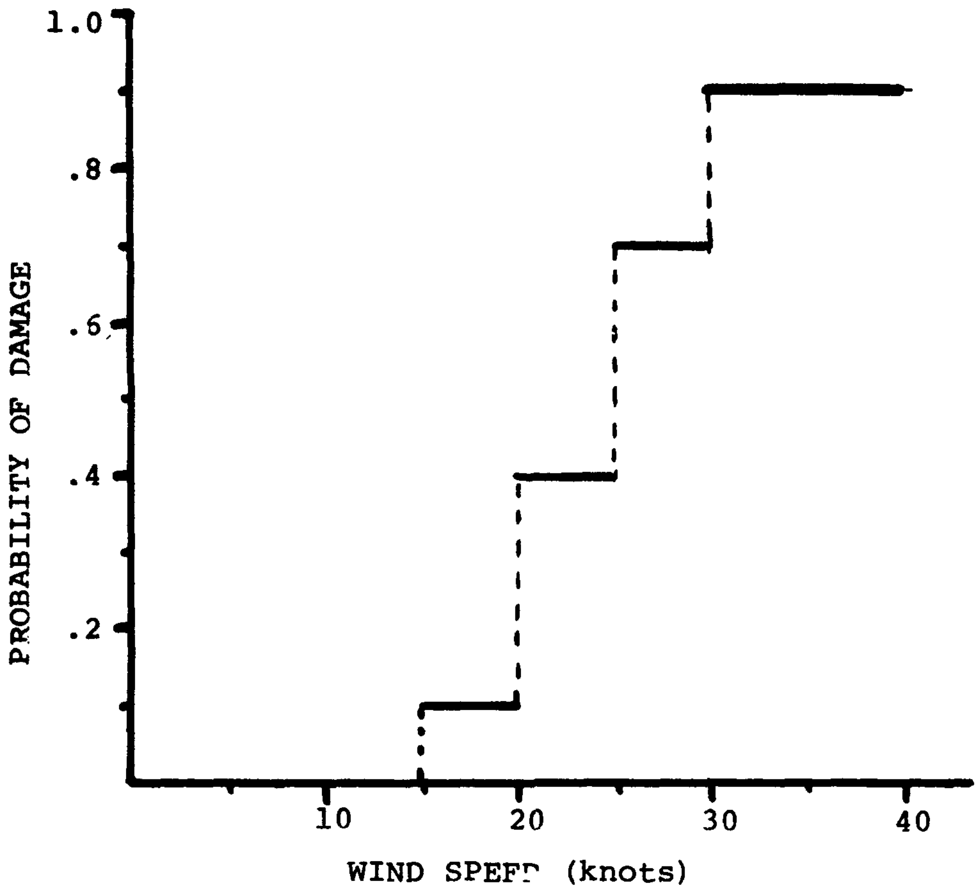


Figure 6-9. Conditional probability of equipment damage given wind speed for paradrop.

intervals, we must obtain the probability of wind speeds in the same intervals. These probabilities are obtained from Figure 6-1 using the differences in cumulative probabilities as shown below:

Probability of speed ≤15 knots = 40%

Probability of 15< speed ≤20 knots = 70% - 40% = 30%,

Probability of 20< speed ≤25 knots = 85% - 70% = 15%,

Probability of 25< speed ≤30 knots = 95% - 85% = 10%,

Probability of speed >30 knots = 100% - 95% = 5%.

These probabilities are probability densities for each speed interval.

The probability of the radio being damaged is found in each interval just as it was for the simple threshold model by multiplying the conditional probability for damage for a wind speed interval by the probability of winds in that interval. Thus, we obtain:

Probability of speed ≤15 kts x Conditional probability of damage (speed ≤15 kts) = 0.4 x 0.0 = 0.0, or 0%,

Probability of 15<speed ≤20 kts x Conditional probability of damage (15<speed ≤20 kts) = 0.3 x 0.1 = 0.03, or 3%,

Probability of 20<speed ≤25 kts x Conditional probability of damage (20<speed ≤25 kts) = 0.15 x 0.4 = 0.06, or 6%,

Probability of 25<speed ≤30 kts x Conditional probability of damage (25<speed ≤30 kts) = 0.1 x 0.7 = 0.07, or 7%,

Probability of speed >30 kts x Conditional probability of damage (speed >30 kts) = 0.05 x 0.9 = 0.045, or 4.5%.

The probability of damage to the radio is the summation of the probability of damage for all speed intervals, or 0.00 + 0.03 + 0.06 + 0.07 + 0.045 = 0.205, or 20.5%. This leads to an MSI of 1.0 - .205 = .795, or 79.5%, under these conditions.

(2) **Tactical photo reconnaissance example.** An RF-4 flying at 20,000 feet requires a cloud-free environment between the aircraft and the target area for successful photography. One operator may define the critical weather threshold as 2/8 total cloud cover below 20,000 feet, i.e., a simple threshold model assuming mission failure if the threshold is exceeded. If the operator receives a categorical forecast of more than 2/8 cloud cover below 20,000 feet, he must either cancel the mission or ignore the forecast. Ideally, the decision maker should know the likelihood of favorable weather so that he may weigh the chance of success against other factors.

There is no guarantee of success with 2/8, or less, total cloud cover. The only cloud in the sky might be right over the target. On the other hand, a break in an almost complete overcast may be over the target, allowing successful photography. Considering this, the operator might better define the probability of successful photography as a function of cloud cover as in Figure 6-10.

The WII needed to support this weather effect model is a continuous probability distribution for cloud cover below 20,000 feet. The forecast distribution could be determined from individual forecasts for each eighth of cloud cover (with the constraint that the sum be exactly 1.0), or determined from a cumulative probability distribution formulated in the manner shown in Attachment 10. A forecast probability distribution is shown in Table 6-4.

The WII shown here is effectively a probability density function, which can be directly multiplied by the conditional probability of "seeing" the target to obtain the probability of successful photography. The calculations are indicated in Table 6-4, with a resultant MSI of 68%.

Note that the most probable coverage is 4/8 (30%). A categorical forecast of this would be a "no go" for a threshold of 2/8. A probability forecast for 2/8 or less coverage would lead to an MSI of 30% (the sum of probabilites for 0 - 2/8 cloud cover). If the critical MSI for proceeding with the mission is between 30% and 68%, a simple threshold forecast will result in cancellation while the more realistic continuous model indicates the mission should be executed.

(3) **Airborne operation example.** Routine paratroop training jumps are only conducted when the drop zone winds are less than 13 knots to minimize the risk of injury. As speed increases, the probability of injury increases dramatically, approaching 100% at some high wind speed. The conditional probability of landing uninjured versus wind speed can be represented by a continuous curve like the one in Figure 6-11.

An airborne unit is given a mission to disrupt enemy communications behind enemy lines and capture key supply and transportation points. It is estimated that a thousand men will be needed on the ground to accomplish this. The importance of the mission is such that it must go at a given time, even if the winds are unfavorable.

The forecaster predicts the wind speed in the drop zone will be about 15 knots. After careful assessment of the weather situation, he derives the forecast cumulative probability distribution of the wind speed using the method of Attachment 10. The forecaster predicts no chance for calm winds, and probabilities of 12.5, 25, 50, 75, 87.5, and 100% for wind speeds below 9, 12, 15, 18, 20, and 35 knots, respectively. These values are plotted in Figure 6-11, and a smooth curve drawn through them to complete the forecast distribution.

The forecast probability density distribution for the wind speed is also shown in Figure 6-11. This distribution was determined by graphically differentiating the cumulative probability curve.

The MSI, based only on wind speed, is the integral of the product of the forecast wind speed probability density (P) and the conditional probability of landing uninjured given the wind speed (PU) over all possible values of wind speed. This integral is shown on Figure 6-11. Since analytic expressions for P and PU are not normally available, we must do the integration by summation as we did in the previous example. These results are presented in Table 6-5, with an MSI of 87%.

Given an MSI of 87%, or 87% probability of landing uninjured, we need to calculate how many paratroopers must be committed to the operation to give an expected force on the ground which is as large as the required 1000 men. Dividing 1000 by .87, we find that 1150 men

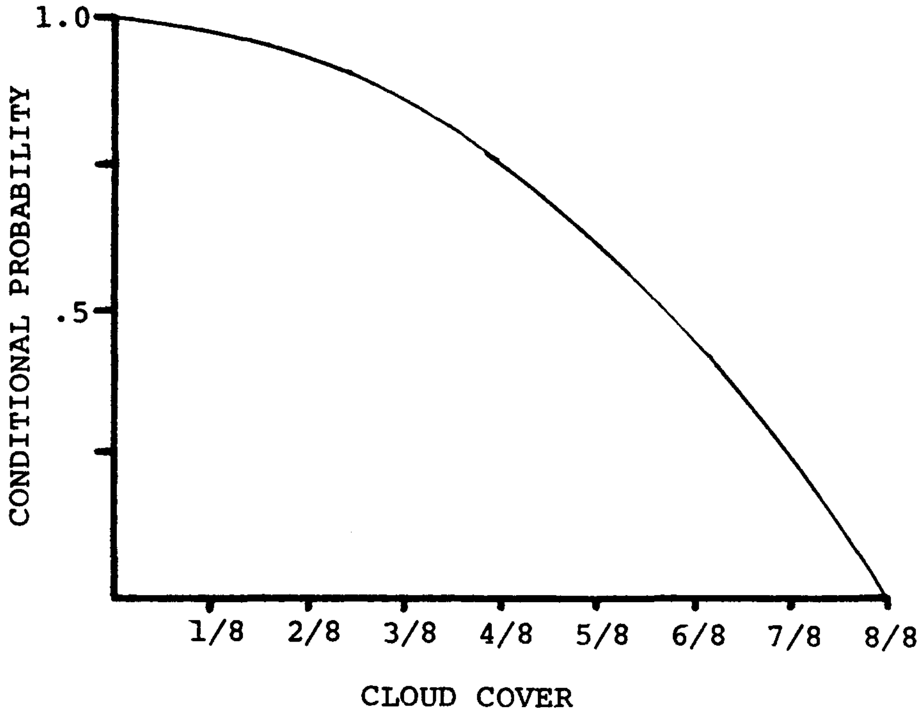


Figure 6-10. Conditional probability of a cloud-free line of sight from 20,000 feet to the surface, given the cloud cover below 20,000 feet.

Table 6-4

Cloud Amount (w)	0/8	1/8	2/8	3/8	4/8	5/8	6/8	7/9	8/8
Forecast Probability of occurrence P(w)	0.05	0.10	0.15	0.20	0.30	0.15	0.05	0.00	0.00
Conditional probability of success C(w) (from Figure 6-10)	1.00	0.95	0.90	0.80	0.65	0.50	0.35	0.20	0.00
Joint Probability P(w) x C(w)	0.05	0.095	0.135	0.160	0.195	0.075	0.018	0.00	0.00

$$\text{Probability of successful photography} = \sum_{w=0}^{8} P(w) \times C(w)$$

$$= 0.05 + 0.095 + 0.135 + 0.160 + 0.195 + 0.075 + 0.018 + 0.00 + 0.00$$

$$= 0.683 = 68\%$$

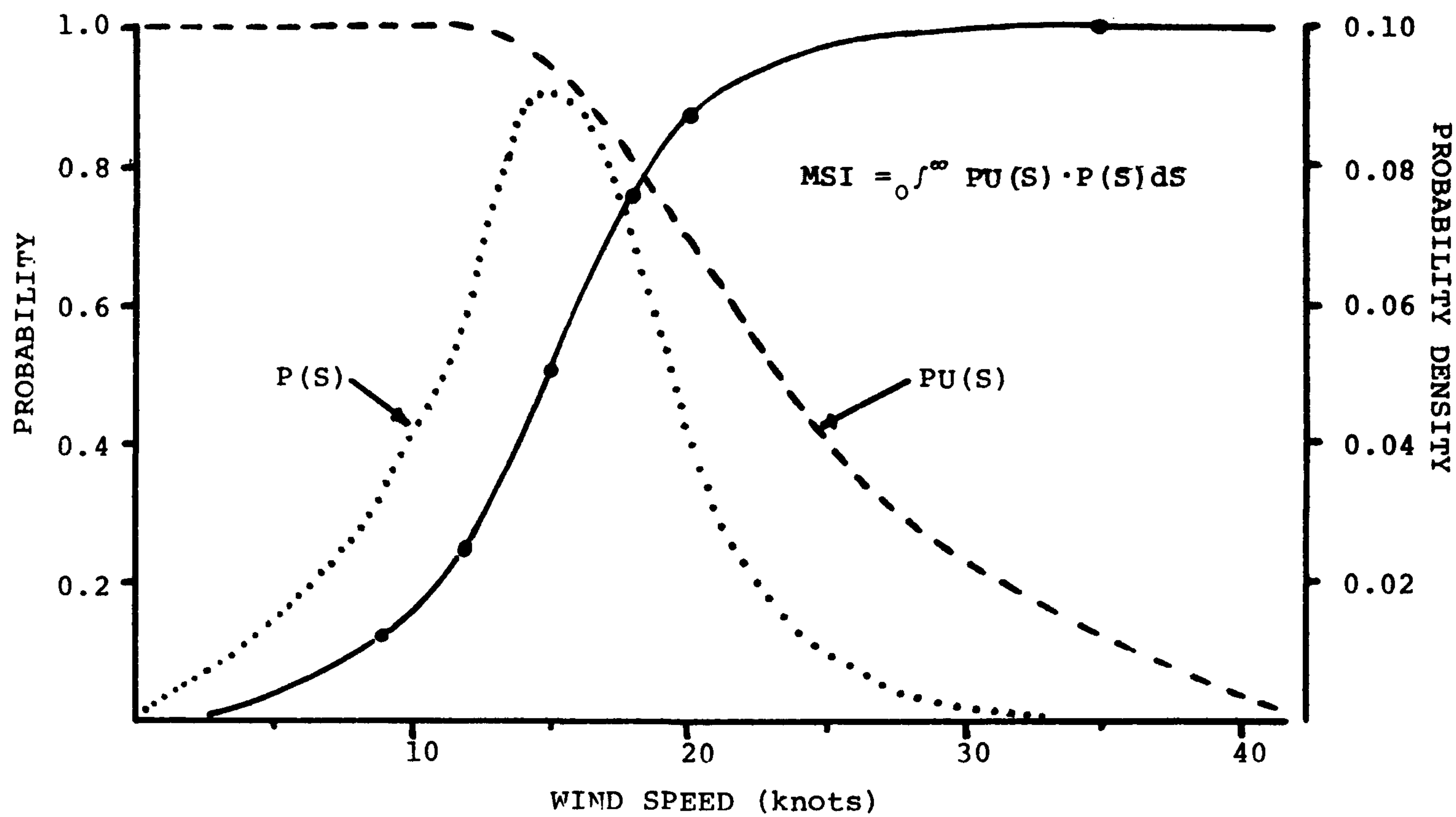


Figure 6-11. Probability curves for the airborne operation example. The solid curve is the cumulative probability for wind speed. The dashed line is the conditional probability of landing uninjured, given wind speed (S). The dotted line is the probability density curve for wind speed.

Table 6-5

Wind Speed Interval	Forecast Cum Prob	Forecast Prob	Cond Prob Landing Uninj	Fract Uninj
0-5	0.035	0.035	1.00	0.035
5-10	0.155	0.120	1.00	0.120
10-12.5	0.280	0.125	1.00	0.125
12.5-15	0.500	0.220	0.97	0.213
15-17.5	0.715	0.215	0.89	0.191
17.5-20	0.875	0.160	0.76	0.122
20-22.5	0.945	0.070	0.61	0.043
22.5-25	0.975	0.030	0.46	0.014
25-30	0.995	0.020	0.31	0.006
30-35	1.000	0.005	0.18	0.001
		1.000		0.870

are needed. Allowing for some margin of error, and to provide care for the injured, the commander may decide to allocate 1200 paratroopers for the mission.

(4) **Weapon selection example.** A different MSI may be computed for each decision option of a single mission. An example of this occurs when several targets are available, and a variety of tactics and weapons, each with a different sensitivity to weather, could be used against each one. In this situation, the probability of success for each weapon type and tactic, at each target, must be calculated.

A simple example of the possible variations of weather impact on the destruction of five targets is given in Table 6-6. The numbers are WIIs tailored to the weather sensitivity of each munition type and delivery mode when a simple threshold weather effect model is assumed. Considering only the weather effect on success, these are also partial MSIs. (If weather is the only factor that affects the mission success, then these are complete MSIs.)

With no other considerations, the weapon and delivery mode selection for each target is straightforward. Simply pick the combination that provides the highest probability of success, provided that probability is greater than the critical probability for flying the mission. For example, if only conventional (visual) weapons are to be used, and 50% is the critical probability, the mission would be flown against E-24, J-14, and K-7, using a low delivery mode. No mission would be flown against E-22 or K-27 because their MSI is less than the critical value.

b. **Non-weather factors.** Obviously, weather is only one of many factors that govern the decisions made concerning a mission. These other factors affect the critical probability required to execute the mission; they affect the actual MSI for the mission; and they may cause decisions to be made that are contradictive to the MSIs involved.

If the radio in the first example is damaged, can it be replaced? How soon can it be replaced? Exactly how important is the radio for the conduct of further operations? These questions all affect what the critical probability of success is for delivery of the radio. What hostile actions may be expected? How large is the drop zone? What is the terrain? These factors all affect the MSI. What are the MSIs for other available delivery modes—helicopter, ground vehicle, land a C-130, etc.? This may cause another mode of delivery to be used.

In the reconnaissance example, enemy defenses affect the loiter time and the probability of success. The importance of the mission governs the critical probability at which the mission will be attempted.

In weapons selection, the number of weapons of different types on hand affects the decision of which type to use. The MSI for one weapon type may be very high against a certain target, but the high unit cost may limit the use of it, except in special circumstances.

What if there are more targets than there are sorties available? If they are all of equal importance, strike the ones with the highest MSI. If they are not of equal importance, their relative value must be considered in making the decision. Suppose relative target values were assigned in the weapon selection example of 2 for E-22, 1 for E-24, 0.5 for J-14, 1.4 for K-7, and 3 for K-27, where the larger values indicate greater importance. These values can be used to weight the MSIs to select which targets to strike. If we multiply the respective MSIs by the weight, we get 0.9 for E-22, 0.75 for E-24, 0.4 for J-14, 0.7 for K-7, and 0.75 for K-27. Now if only three sorties were available, they should be flown against E-22, E-24, and K-27. Note that a mission is not flown against J-14, the one with the highest MSI, because of it's relative importance.

The commander has many factors to consider when trying to arrive at an optimal decision. A categorical forecast must be interpreted in order to assess the true impact of weather. Forecasters who work closely with their customer may attempt to adjust their categorical forecasts, according to their understanding of the critical probability, in effect making mission decisions without knowing all the facts. The WII eliminates the need to interpret the forecast, allowing the commander, whose job it is to know and assess all mission factors, to make the best use of weather in planning and execution.

6-5. Categories of WII. There are three categories of WII: forecast, climatological, and simulated. Each is designed for a specific purpose.

a. **Forecast WII (FWII).** The examples of WIIs which have been presented in this chapter are of this type. FWIIs are normally used in the execution phase of a mission. They are also used for short range planning. when forecasts would be expected to have more skill than climatology. FWIIs are produced in several ways.

(1) Centralized facilities produce generalized categorical weather products and guidance—surface, upper air, and HWD analyses and progs, etc. Local forecasters and SWOs combine these with more recent observations and climatological aids to produce subjective threshold forecasts and probability distributions.

(2) General probability forecasts from weather centrals—MOS bulletins, area forecasts, etc—are tailored to specific customer needs by local forecasters and SWOs.

(3) FWIIs produced by weather centrals are modified, as required, by local forecasters and SWOs before providing them to decision makers.

(4) A weather central produces FWIIs and perhaps even partial MSIs if the customer requests them (such as advanced CFPs, TERBs, etc), and transmits them directly to the decision maker in a tailored bulletin.

(5) Forecast probability distributions of weather elements are transmitted from a weather central to a customer's computer, where they are used in the production of MSIs. Advantages of this method are: a great reduction in communication volume: the weather information is unclassified; and complete, automated tailoring with climatology, targets, times, and non-weather factors to produce MSI. Within his own computer the customer can fully incorporate weather impacts to produce MSIs for all possible options; play as many "What if?" games as he wishes, frag aircraft; plan missions; and anticipate weather constraints on enemy operations. Circumstances and the sophistication of customer applications will dictate the method used to produce FWIIs.

b. Climatological WII (CWII). Much of our climatological information is already in probabilistic form. Tailored climatological probabilities are routinely provided to customers for planning, scheduling, selecting areas and routes. SOCS/RUSSWO data is ideal

Table 6-6. Weather Impact Indicators for use in weapon selection.

TARGET	DELIVERY	BOMB TYPE				
		VISUAL	TV	INFRARED	MICROWAVE	LASER
E-22	Hi	30	25	30	85	20
	Low	45	40	50	85	35
E-24	Hi	40	35	80	85	30
	Low	75	70	75	85	75
J-14	Hi	30	30	30	45	25
	Low	80	75	80	85	85
K-7	Hi	35	30	25	40	25
	Low	50	45	40	55	40
K-27	Hi	10	10	05	15	05
	Low	25	20	25	35	20

for generating probabilities for simple thresholds or probability distributions for continuous/multiple thresholds.

c. Simulated WII (SWII). SWIIs are used to show the expected effect of weather on mission accomplishment, attrition, and resource requirements. SWIIs can be used by a customer to simulate MSIs, and thus help determine the desirability of various force structures, weapon systems, tactics, and force distributions. SWIIs are produced by a model which simulates the variability of observed weather for a climatic regime and the accuracy of weather forecasts. Known time and space correlations of generated observations are included in the model. The time decay of forecasting skill is taken into account. SWIIs allow comparison of mission results based on, say, 12 hour forecasts with those using 6-hour forecasts.

One type of SWII can be used to help the customer determine the critical probability for go/no-go decisions. Critical probability can be determined objectively if the relative utilities of the various mission outcomes are known (see section 5-6). However, this is rarely the case. But SWIIs help the decisionmaker use his "gut feelings" on the desirability of mission outcomes to select his critical probability. An example will show how SWIIs meet this purpose.

Suppose a customer needs a 12-hour probability forecast for a critical weather threshold[1]. Climatological records show this threshold is exceeded 40% of the time.

(1) What is the expected distribution of probability forecasts for this event? A forecaster making two-week forecasts for this event would always predict a probability of 40%, the climatological frequency. A forecaster making two-minute probability forecasts would predict 0% probability nearly 60% of the time (the threshold is not exceeded now). He would predict 100% probability nearly 40% of the time (the threshold is exceeded now). He would predict some intermediate probability a very small fraction of the time (the weather is very close to the threshold now, and it could go either way in two minutes). Two weeks in advance, the forecaster would almost never forecast a 0% or 100% probability for exceeding the threshold. Two minutes in advance, the forecaster would rarely be so uncertain that he would issue a 40% probability forecast. Between these two extremes the relative frequencies of the various probabilities given by the forecaster should differ, depending on the length of the forecast. This change in forecast frequency distribution with the length of the forecast is shown in Figure 4-2. The second row in this figure illustrates the frequency distribution for probability forecasts for a threshold with a 40% climatological expectation. The distribution in the leftmost column (0.2 correlation) of Figure 4-2 is about equivalent to a three-day forecast. The distribution in the rightmost column (0.95 correlation) of Figure 4-2 is nearly that expected for a three-hour forecast. The distribution for twelve-hour forecasts is close to that shown in the fourth column (0.8 correlation). The distribution for twelve-hour forecasts of a threshold with a 0.4 climatological frequency is shown in the original, continuous form as the curve labeled ψ in Figure 6-12. This distribution is based on an application of the Transnormalized Regression Probability Model. If a sufficient record is available, actual probability forecast distributions for the threshold could also be used, after some subjective smoothing and adjustments for possible sampling error using Figure 4-2 as a guide.

(2) The frequency distribution takes the sharpness (skill) of the probability forecasts into account. The reliability of the probability forecasts must also be included in the formulation of SWIIs. The assumption made is that the forecasts are perfectly reliable. Actual reliability experience could be used. However, forecasters can learn to make reliable forecasts with practice, and try to eliminate personal biases. Deviations from perfect reliability in one sample of forecasts may be in the opposite direction for a subsequent sample. Thus, perfect reliability is usually the best assumption. Perfect reliability is indicated by the curve θ in Figure 6-12. The ratio θ/ψ indicates the fraction of the forecasts at each probability in which the threshold is exceeded. θ is zero when the forecast probability is zero—the threshold is never exceeded, when the forecast probability is zero. At 100% forecast probability the ψ and θ curves have the same value—the threshold is always exceeded, when the forecast probability is 100%. At intermediate probabilities, the θ curve lies at a distance from the horizontal axis to the ψ curve, proportionate to the forecast probability. For example, at 40% forecast probability, the θ curve is 40% of the value of the ψ curve. The θ curve has a value 75% of the ψ curve at a 75% forecast probability. Thus the θ curve represents perfect reliability for the frequency distribution of the probability forecasts.

(3) The θ curve separates the occasions when the threshold is exceeded from those when it is not exceeded. The area between the horizontal axis and the curve is the portion of the occasions (for all forecasts) in which the threshold is exceeded. This area is 40% (the climatological frequency) of the total area under the ψ curve in this case. The area between the θ and ψ curves is the portion of the occasions for all forecasts, 60%, when the threshold is not exceeded. Suppose that a critical probability value of 40% is selected. This is represented by the dashed vertical line in Figure 6-12. The customer will always execute, if the forecast probability exceeds this value. The entire area between the curve and the horizontal axis to the left of the critical probability line is the portion of the total number of forecasts that will be less than the critical probability. To the right of the critical probability line, it is the portion of forecasts greater than the critical probability. Together, the ψ and θ curves and the critical probability

[1]The discussion will address the probability for exceeding a given weather threshold - 1000′ ceiling, 3 miles visibility, etc. It applies equally to the occurrence/non-occurrence of a yes/no event, e.g., rainfall, thunderstorm.

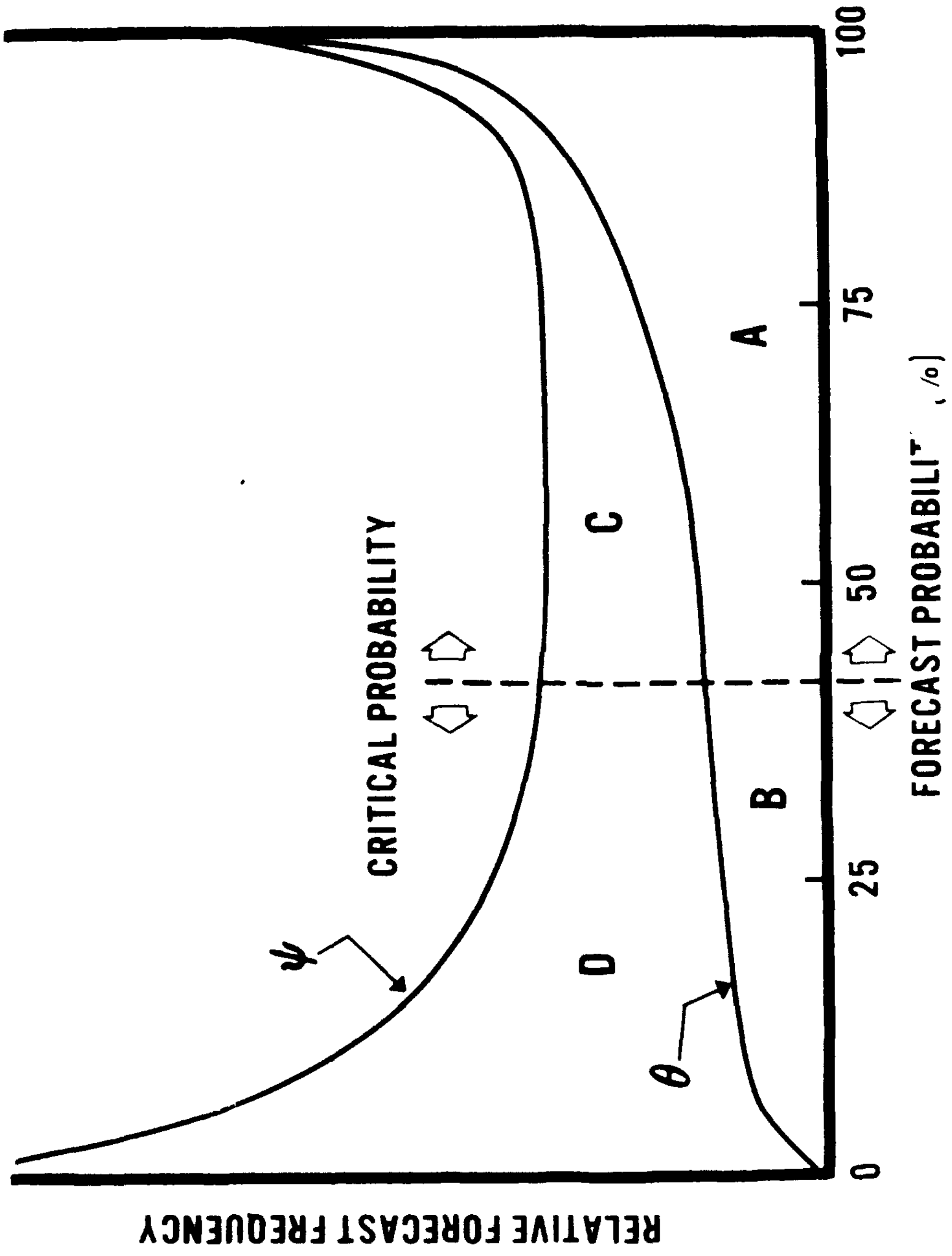


Figure 6-12. Effect of Critical Probability on Operational Verification.

line separate the forecasts into four outcomes, labeled A, B, C, and D in Figure 6-12. Area A represents the portion of the time the forecast probability will be greater than the critical probability, and the threshold is exceeded, these represent correct "go" forecasts. Area C is the portion of time the forecast probability will be greater than the critical probability, but the threshold is not exceeded. These are scheduled/attempted mission executions that will have unfavorable weather (i.e., aborts). Missed opportunities, mission stand-downs with observed favorable weather, are given by area B. Area D is the fraction of correctly cancelled missions, those with subsequently observed unfavorable weather. The areas are part of the resultant two-by-two verification matrix, when the critical probability is used to make go/no-go decisions. This matrix is shown in Table 6-7.

(4) The ratios of the four areas—A, B, C, and D—to the total area under the Ψ curve give the fractions of the time respectively that the customer would: expect to execute a mission with favorable weather; not execute and have favorable weather; mission aborts/cancellations due to unfavorable weather; and correct standdowns because of weather. Remember, these outcomes are those expected for an event with a 40% climatological frequency, using a twelve-hour probability forecast and a selected critical probability for mission execution. Different critical probabilities will change the proportions of the mission outcomes. If the dashed vertical line for critical probability in Figure 6-11 were moved left or right, the relative size of areas A, B, C, and D would change. A critical probability of 0% (execute regardless of the weather forecast) would reduce areas B and D to zero and enlarge areas A and C to 40 and 60% of the total area, respectively. The user would expect the climatological frequency of favorable and unfavorable weather at mission execution. At the other extreme, a critical probability of 100%, never go, would result in a 40% frequency of missed opportunities, and a 60% rate of correct stand-downs. This variation in the mission outcomes with critical probability is shown in Figure 6-13.

(5) A decision maker can use graphical aids like Figure 6-13 to adjust his critical probability to obtain the desired rate of missed opportunities, false alarms, prefigurance, postagreement, etc. One who wanted to minimize missed opportunities would select an appropriately low critical probability. Another who needed to execute against a well defended, fixed target might select a high critical probability that would minimize C, the mission abort rate due to weather, and thus the unnecessary exposure of aircraft to hostile fire. USAFETAC can produce graphs like Figure 6-13 for various elements, thresholds, and forecast lead times. An example is shown in Table 6-8. The columns labeled A, B, C, and D in this table identify the relative frequencies for the corresponding matrix positions of Table 6-7 for the given critical probabilities.

(6) USAFETAC calculated a series of SWII tables similar to Table 6-8 using the method described in this section. The tables cover a large number of event/threshold climatological frequencies and forecast skills (correlations). SOCS or other climatic aids can be used to determine the frequency of the event/threshold for the desired time of day and year. The correlation for predicting ceiling and visibility thresholds can be estimated by $R = .98^{t}$, where t is in hours.[1] If a history of categorical or go/no-go forecast verification for the event/threshold is available, the correlation between forecasts and observations for the sample can be calculated using the tetrachoric correlation formula in AWS TR 75-259, page 23.

[1] This form is derived from correlation values given for extratropical regions in Touart, 1973.

Table 6-7. Verification Matrix for Critical Probability. (Assuming weather is the only factor in mission success.)

OBSERVED	FORECAST GO	FORECAST NO GO
GO	A	B
NO GO	C	D

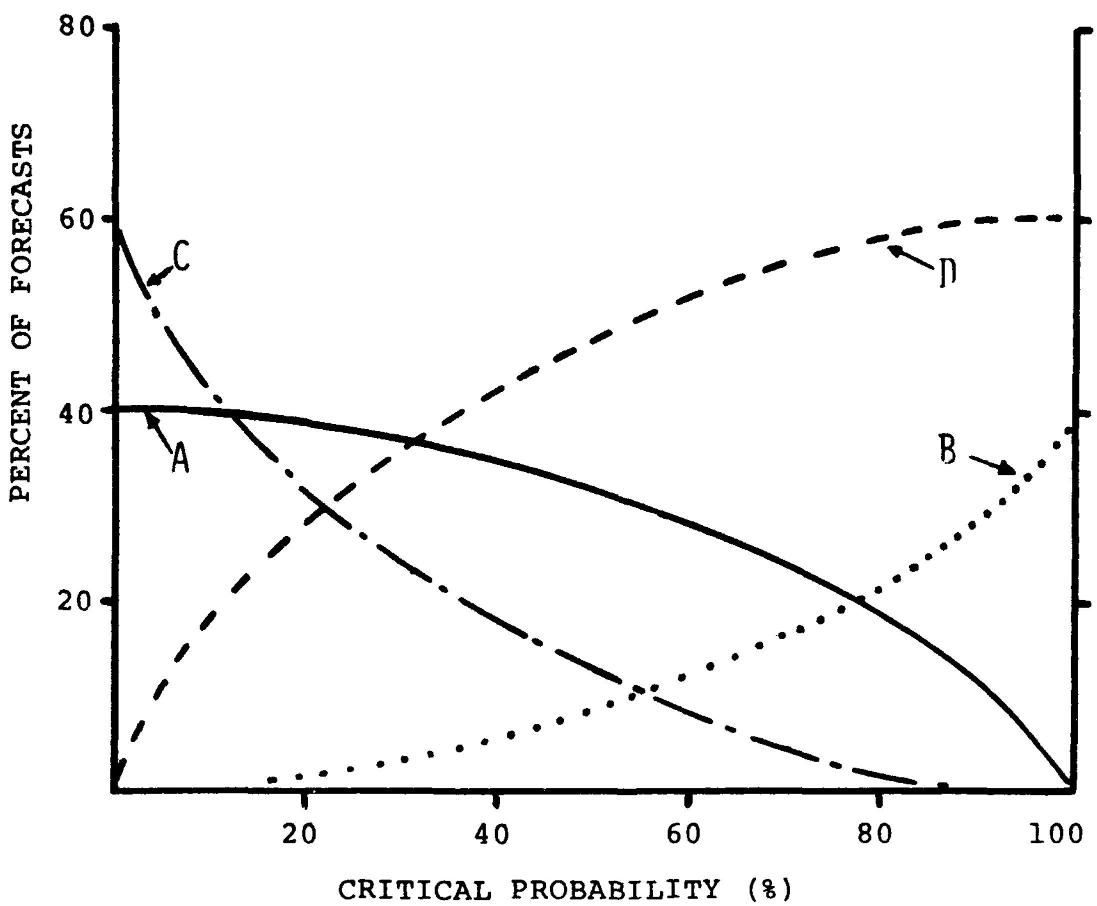


Figure 6-13. Sample of graphical presentation of data such as is in Table 6-8.

USAF ENVIRONMENTAL TECHNICAL APPLICATIONS CENTER (USAFETAC) SIMULATED WEATHER IMPACT INDICATORS- 30-MAR-77

STATION : TRAVIS AFB CAL MONTH : NOV CLIM PROB : 0.890

CATEGORY : >300/.75 START TIME : 01L FORECAST TIME LENGTH = 6 HRS ** VALID AT : 07L

CRITICAL PROBABILITY	MISSION EXEC WITH SUCCESS**	MISSION NOT EXEC WOULD HAVE SUCCEEDED	MISSION EXEC DID NOT SUCCEED	MISSION NOT EXEC AND WOULD NOT HAVE SUCCEEDED
0.050	0.890	0.000	0.098	0.012
0.100	0.889	0.001	0.091	0.019
0.150	0.888	0.002	0.085	0.025
0.200	0.887	0.003	0.079	0.031
0.250	0.886	0.004	0.074	0.036
0.300	0.883	0.007	0.068	0.042
0.350	0.881	0.009	0.063	0.047
0.400	0.878	0.012	0.058	0.052
0.450	0.874	0.016	0.053	0.057
0.500	0.869	0.021	0.047	0.063
0.550	0.864	0.026	0.042	0.068
0.600	0.857	0.033	0.037	0.073
0.650	0.849	0.041	0.033	0.077
0.700	0.839	0.051	0.028	0.082
0.750	0.826	0.064	0.023	0.087
0.800	0.809	0.081	0.018	0.092
0.850	0.787	0.103	0.013	0.097
0.900	0.754	0.136	0.009	0.101
0.950	0.696	0.194	0.004	0.106

** Mission success or failure is based solely on weather factors.

Table 6-8. Example of Automated SWIIs Available from USAFETAC.

The above example applies to a system sensitive to a simple threshold. Equivalent techniques are available for continuous threshold models. SWIIs for the systems using continuous threshold models can be requested from USAFETAC.

Chapter 7

Implementation of Probability Forecasts

7-1. General. The success of a probability forecast program depends to a great degree on how it is implemented. This chapter recommends how to implement a probability forecast program at the detachment level. For most applications the program should evolve through four phases: development, testing, evaluation, and operational use.

7-2. Development. Choosing and defining the forecast event is the first and most critical step. Weather events with the **most operational** impact should be chosen first. This step requires very close coordination with the customer to precisely define an event which is operationally significant and within forecasting capability. Tailored threshold forecasts should be considered first for most requirements. If there are several customers with similar requirements, consider a more general forecast. Do not attempt to furnish weather impact indicators until the unit has thoroughly **mastered probability forecasts, and the customer** understands how to use them. Since the customer's ability to use probabilities is just as important as the quality of the forecasts, he should understand the decision models, critical probabilities, and other procedures used in the decision process. The detco or SWO should take the leading role in identifying where probability forecasts can be applied and advising the user. Contact the parent squadron or wing consultant if outside assistance is needed.

7-3. Testing. This step determines the feasibility of satisfying the user's requirement. Once an event is defined, a test is needed to evaluate if the forecasts meet customer requirements. The SWO must coordinate with the customer to establish the standards of acceptable reliability for the operation under consideration. A straight 5% deviation from perfect reliability (bias) might be used. A deviation of 5% of the forecast probability (i.e. 5% bias at 100% probability, a 2.5% bias at 50% probability, 0.5% bias at 10% probability, etc) might be more appropriate, especially when the customer's critical probability is very low and very sensitive. The customer should be shown reliability diagrams depicting upper and lower limits, so he will know the limits of your capability.

7-4. Evaluation. Evaluation is a continuing process, but always of more importance initially. Forecasters inexperienced in probability forecasting must be trained. All forecasters must be trained when a new forecasting requirement (event) is undertaken. Attachment 8 has a training scenario that can be used. Both types of training (new forecasters and new events) are necessary to establish reliability in the forecasts. This should be done prior to going operational. After the forecasts are implemented, feedback of the reliability of the forecasts should be provided to the customers on a periodic basis. Forecasters should be provided frequent feedback on the reliability of their forecasts, so they may gain experience in quantifying uncertainty.

7-5. Operational Use. Implementation should not be rushed. The unit should be thoroughly prepared to issue probability forecasts, and the customer fully knowledgeable on how to use them properly. This is especially true for the first attempts. If things go wrong, the customer will undoubtedly be reluctant to further use them. It is also important that the customer know that the payoff from using probability forecasts is cumulative, and can only be realized if these forecasts are used consistently over an extended period.

HQ AIR WEATHER SERVICE
OFFICIAL
U.S. AIR FORCE

ALBERT J. KAEHN, JR, *Colonel, USAF*
Commander

CHARLES W. TIEMANN
Administration Officer

DISTRIBUTION F;X
MAC/IGIS 1
Det 6, HQ AWS 45
Naval Oceanography Command
National Space Technology Laboratories
Bay St. Louis, MI 39520 70
Naval Environmental Prediction Research Facility
Attention: Library
Monterey, CA 93940 4
US Department of Commerce
National Oceanic and Atmospheric Administration
National Weather Service/W111
8060 13th Street
Silver Spring, MD 20910 100

BIBLIOGRAPHY

Following is a list of the most useful references consulted in preparing this pamphlet. References preceded by an asterisk are recommended for retention in unit technical libraries as a supplement to this pamphlet. Relevant information in other references has been embodied in this pamphlet. This list is included to simplify literature searches for those desiring additional background information.

Adams, R. L.A., 1974: The differential use of personal observations and weather forecasts in making New England beach trip decision. Preprints, Fifth Conference on Weather Forecasting and Analysis (St Louis). Boston, Mass., American Meteorological Society, pp 40-43.

*AWSTR 77-267: Guide for Applied Climatology, Scott AFB IL, 161 pp.

*AWSTR 75-259: Transnormalization Regression Probability. Scott AFB IL.

*AWSTR 77-273: Selected Topics in Statistical Meteorology. Scott AFB IL.

Bocchieri, J. R., R. L. Crisci, H. R. Glahn, F. Lewis, and F. T. Globokar, 1974: Recent developments in automated prediction of ceiling and visibility.

J. Appl. Meteor., 13, 277-288

Bocchieri, J. R. and H. R. Glahn, 1972: Use of model output statistics for predicting ceiling height. Mon. Wea. Rev., 100, 869-879.

Boehm, A. R., 1973: Estimating utility of AWS forecasts using product evaluation program (PEP) categories. Scott AFB IL., USAF Environmental Technical Applications Center, Aerospace Sciences Review, 73-1, 10-17.

Boehm, A. R., 1976a: How to use AFGWC point warning probabilities. Scott AFB IL., USAF Environmental Technical Applications Center, Aerospace Sciences Review, 76-1, 45-58.

Boehm, A. R., 1976b: A priori estimates of distributions of probability forecasts and relationships to the value of forecasts. Weather Forecasting and Weather Forecasts: Models, Systems, and Users, National Center for Atmospheric Research, Boulder CO.; Vol 2, 854-871.

Boehm, A. R., 1976c: Optimal decisions through mission success indicators. Proceedings of the 7th Technical Exchange Conference, (El Paso), Atmospheric Sciences Laboratory, White Sands Missile Range, NM, pp 17-25.

Bosart, L. F., 1975: SUNYA experimental results in forecasting daily temperature and precipitation. Mon. Wea. Rev., 103, 1013-1020.

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

Bross, I. D. J., 1953: Design for Decision. New York, The Free Press. (Paperback-$2.45). 276 pp.

Charba, J. P., 1975: Operational scheme for short range forecasts of severe local weather. Preprints, Ninth Conference on Severe Local Storms (Norman). Boston, Mass., American Meteorological Society, pp. 51-57.

Christensen, R. E., 1975: Use of operational decision models. Scott AFB IL, USAF Environmental Technical Applications Center, Aerospace Sciences Review, 75-4, 7-12.

Cummings, R. L., 1969: Meteorological decision making and Bayesian analysis. Scott AFB IL, USAF Environmental Technical Applications Center, Aerospace Sciences Review, 69-1, 14-19.

Epstein, E. S., 1962: A Bayesian approach to decision making in applied meteorology. J. Appl. Meteor., 1, 169-177

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. J. Appl. Meteor., 8, 985-987.

Epstein, E. S. and R. J. Fleming, 1971: Depicting stochastic dynamic forecasts. J. Atmos Sci., 28, 500-511.

Epstein, E. S., and A. H. Murphy, 1965: A note on the attributes of probability predictions and the probability score. J. Appl. Meteor., 4, 297-299.

Glahn, H. R., 1964: The use of decision theory in meteorology. Mon. Wea. Rev., 92, 383-388.

Glahn, H. R., 1974: Problems in the use of probability forecasts. Preprints Fifth Conference on Weather Forecasting and Analysis. Boston, Mass., American Meteorological Society, pp 32-35.

Glahn, H. R., 1975: Making probability forecasts by objective methods. Scott AFB IL, USAF Environmental Technical Applications Center, Aerospace Sciences Review, 75-3, 13-23.

Glahn, H. R., and J. R. Bocchieri, 1975: Objective estimation of the conditional probability of frozen precipitation. Mon. Wea. Rev., 103, 3-15.

Glahn, H. R. and D. L. Jorgensen, 1970: Climatological aspects of the Brier P-score. Mon. Wea. Rev., 98, 136-141.

Glahn, H. R., and W. H. Klein, 1975: Present status of statistical weather forecasting. Pre-prints, Fourth Conference on Probability and Statistics in Atmospheric Sciences (Tallahassee). Boston, Mass., American Meteorological Society, pp, 1-6.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. J. Appl. Meteor., 11, 1203-1211.

Gleeson, T. A., 1960: A prediction and decision method for applied meteorology and climatology, based partly on the theory of games. J. Meteor., 17, 116-121.

Globokar, F. T., 1976: Probability Weather Forecasts: A Viable Alternative. Unpublished research paper, Air Command and Staff College, Air University, Maxwell AFB AL. Report No. 1025-76, 113 pp.

Gringorten, I.I., 1972: Conditional Probability for an Exact, Noncategorized Initial Condition. Mon. Wea. Rev., 100, 796-798.

Gringorten, I. I., 1967: Verification to determine and measure forecasting skill. J. Appl. Meteor., 6, 742-747.

Grubbs, B. E. and T. C. Wann, 1975: AFGWC Verification of MOS Generated Forecasts for Selected AF Stations. Unpublished paper, AFGWC, Offutt AFB NE, 16 pp.

Hall, D. P., 1972: An operational decision model employing operational and environmental factors. Scott AFB IL., USAF Environmental Technical Applications Center, TN 72-8, 8 pp.

Hansen, D.C. 1969: Limiting probabilities as applied to synoptic-scale analysis and prediction. Scott AFB, IL., USAF Environmental Technical Applications Center, Aerospace Sciences Review, 69-1, 7-11.

Hansen, D.C., 1975: Probability of ceiling/visibility on TAFs. Scott AFB IL, USAF Environmental Technical Applications Center, Aerospace Sciences Review, 75-4 18-19.

Hering. W. S., and D. L. Quick, 1974: Hanscom visibility forecasting experiments. Preprints, Fifth Conference on Weather Forecasting and Analysis (St Louis). Boston, Mass., American Meteorological Society, pp 224-227.

Hughes, L. A., 1965: On the probability forecasting of the occurrence of precipitation. ESSA Technical Note 20-CR-3, 36 pp.

Hughes, L. A., 1967a: On the Use and Misuse of the Brier Verification Score. ESSA Technical Memorandum WBTM CR-18, 14 pp.

Hughes, L. A., 1967b: Improving Precipitation Probability Forecast Using the Central Region Verification Printout. ESSA Central Region Technical Memorandum 15, 18 pp.

Hughes, L.A., 1976a: On the verification of forecasts. Preprints, Sixth Conference on Weather Forecasting and Analysis (Albany). Boston, Mass., American Meteorological Society, pp 1-8.

Hughes, L. A., 1976b: Evaluation of Probability Forecasts - Problems and Proposals. Weather Forecasting and Weather Forecasts: Models, Systems, and Users. National Center for Atmospheric Research, Boulder, CO, Vol 2, 839-853.

Julian, P. R. and A. H. Murphy, 1972: Probability and statistics in meteorology: a review of some recent developments. Bull. Amer. Meteor. Soc, 53, 957-965.

Kelly, J. J., 1976: Uncertainty and Weather Forecasts: Must They Remain Mutually Exclusive. Unpublished research paper, Air Command and Staff College, Air University, Maxwell AFB AL, Report No. 1510-76, 99 pp.

Kernan, C. L. 1975: The cost-loss decision model and air pollution forecasting. J. Appl. Meteor., 14, 8-16.

Klein, W. H., and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. Bull. Amer. Meteor. Soc., 55, 1217-1227.

Kolb, L. L. and R. R. Rapp, 1962: The utility of weather forecasts to the raisin industry. J. Appl. Meteor, 1, 8-12.

*Lyon, H. A. and L. L. LeBlanc, 1976: Weather probability forecasts: A cost-saving technique in space launch and missile test operations. Air University Review, Jan-Feb, pp 45-54.

Meyer, W. D., 1975: Probability forecasting for weather-sensitive military operations. Scott AFB IL., USAF Environmental Technical Applications Center, Aerospace Sciences Review, 75-5, 69-81.

Miller, R.G., 1962: Statistical prediction by discriminant analysis. Meteor. Monogr., 4, No. 25, Boston, Mass., American Meteorologial Society, 54 pp.

Miller, R. G., R. C. Whiton, and M. J. Kelly Jr., 1977: Results of a single station forcasting experiment. Preprints, Fifth Conference on Probability and Statistics in Atmospheric Sciences, Las Vegas, American Meteorological Society, Boston, 37-40.

Murphy, A. H., 1966: A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. J. Appl. Meteor., 5, 534-537.

Murphy, A. H., 1969: Measures of the utility of probabilistic predictions in cost-loss ration decision situations in which knowledge of the cost-loss ratio is incomplete. J. Appl Meteor., 8, 863-873.

Murphy, A. H., 1972a: Ordinal relationships between measures of the "accuracy" and "value" of probability forecasts: preliminary results. Tellus, 24, 531-542.

Murphy, A. H., 1972b: Scalar and vector partitions of the probability score: Part I. Two-state situation. J. Appl. Meteor., 11, 273-282.

Murphy, A. H., 1972c: Scalar and vector partitions of the probability score: Part II. N-state situation. J. Appl. Meteor., 11, 1183-1192.

Murphy, A. H., 1973a: Hedging and skill scores for probability forecasts. J. Appl. Meteor., 12, 215-223.

Murphy, A.H., 1973b: A new vector partition of the probability score. J. Appl. Meteor., 12, 595-600.

Murphy, A. H., 1974b: A sample skill score for probability forecasts. Mon Wea. Rev., 102, 48-55.

Murphy, A. H., 1976a: Decision-making models in the cost-loss ration situation and measures of value of probability forecasts. Mon. Wea. Rev., 104, 1058-1065.

Murphy A. H., 1976b: Evaluation of probabilistic forecasts: some procedures and practices. Weather Forecasting and Weather Forecasts: Models, Systems, and Users, National Center for Atmospheric Research, Boulder CO, Vol 2, 807-830.

Murphy, A.H., 1976c: The value of weather forecasts: some theoretical results. Weather Forecasting and Weather Forecasts: Models, Systems, and Users, National Center for Atmospheric Research, Boulder Co, Vol 2, 579-598.

Murphy, A. H., 1976d: Probability forecasting in meteorology: status, problems, and prospects. Preprints, Sixth Conference on Weather Forecasting and Analysis (Albany). Boston, Mass, Amerian Meteorologial Society, pp 9-12.

Murphy, A. H., 1976e: The value of climatological, categorical, and probabilistic forecasts in the cost-loss ration situation. Mon. Wea. Rev., 105, 803-816.

Murphy A. H., 1977: Subjective quantification of uncertainty in real-time weather forecasts in the United States. To appear in Recent Developments in Real-Time Forecasting/Control of Water Resource Systems.

Murphy, A. H. and E. S. Epstein, 1967: Verification of probabilistic prediction systems: a brief review. J. Appl. Meteor., 6, 748-755.

Murphy, A. H., and R. L. Winkler, 1974a: Probability forecasts: a survey of National Weather Service forecasters. Bull. Amer. Meteor. Soc., 55, 1449-1453.

Murphy, A. H., and R. L. Winkler, 1974b: Subjective probability forecasting experiments in meteorology: some preliminary results. Bull. Amer. Meteor. Soc., 55, 1206-1216.

Murphy, A. H., and R. L. Winkler, 1975: Subjective probability forecasting in meteorology: some recent experimental results. Preprints, Fourth Conference on Probability and Statistics in Atmospheric Sciences (Tallahassee). Boston, Mass., American Meteorological Society, pp. 41-48.

Murphy, A. H. and R. L. Winkler, 1976a: Experimental point and area precipitation probability forecasts for a forecast area with significant local effects. To appear in Atmosphere, 15 (1977), No. 2.

Murphy, A. H. and R. L. Winkler, 1976b: The Use of Credible Intervals in Temperature Forecasting: Some Experimental Results. Unpublished paper, 17 pp.

Murphy, A. H. and R. L. Winkler, 1977: Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? National Weather Digest 2, No 2, 2-9.

Myers, J. N., 1974: Limitations of precipitation probability forecasts. Preprints, Fifth Conference on Weather Forecasting and Analysis (St Louis). Boston, Mass., American Meteorologial Society, pp 36-39.

Nelson, R. R. and S. G. Winter, Jr., 1960: Weather Information and Economic Decision, a Preliminary Report. Santa Monica, CA., RAND Corp Research Memorandum RM-2620 - NASA, 121 pp.

Nelson, R. R. and S. G. Winter, Jr., 1964: A case study in the economics of information and coordination, the weather forecasting system. Quart. J. Econ., 78, 420-441.

Panofsky, H.A. and G.W. Brier, 1965: Some Applications of Statistics to Meteorology. Penn State Univ., Univ. Park, Penn, p 199-203. in weather forecasts.

Peterson, C. R., K. J. Snapper, and A. H. Murphy, 1972: Credible interval temperature forecasts. Bull. Amer. Meteor. Soc., 53, 966-970.

Roberts, C.F., 1965: On the use of probability statements in weather forecasts. ESSA Technical Note 8-FCST-1, 15 pp.

Rogell, R. H., 1972: Weather terminology and the general public. Weatherwise, 25, 126-132.

Root, H. E., 1962: Probability statements in weather forecasting. J. Appl. Meteor., 1, 163-168.

Russo, J. A., 1966: The economic impact of weather on the construction industry in the U.S. Bull. Amer. Meteor. Soc., 47, 967-972.

Sanders, F., 1958: The evaluation of subjective probability forecasts. Scientific Report No. 5., Department of Meteorology, MIT, 57 pp.

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

Selvidge, J., 1976: Rapid Screening of Decision Options. Technical Report 76-12, Decisions and Designs, Inc.; McLean, Va. 99 pp.

Smith, D. L., 1977: An examination of probability of precipitation forecasts in light of rainfall areal coverage. National Weather Dig [illegible], No 2, 15-26.

Smith, M. D., 1975: Probability of extreme winds with a passage of a tropical storm. Scott AFB IL, USAF Environmental Technical Applications Center, Aerospace Sciences Review, 75-4, 13-15.

Thompson, J. C., 1950: A numerical method for forecasting rainfall in the Los Angeles area. Mon. Wea. Rev., 78, 113-124.

Thompson, J. C., 1952: On the operational deficiencies in categorical weather forecasts. Bull. Amer. Meteor. Soc., 33, 223-226.

Thompson, J. C., 1962: Economic gains from scientific advances and operational improvements in meteorological prediction. J. Appl. Meteor., 1, 13-17.

Thompson, J. C., 1966: A note on meteorological decision-making. J. Appl. Meteor., 5, 532-533.

Thompson, J. C. and G. W. Brier, 1955: The economic utility of weather forecasts. Mon. Wea. Rev., 83, 249-254.

Touart, C. N., 1973: Synthetic Verification Matrices - I, Air Force Cambridge Research Laboratories Technical Report 73-0158, Air Force Surveys in Geophysics No. 260 (AD-761485).

Winkler, R. L. and A. H. Murphy, 1970: Nonlinear utility and the probability score. J. Appl. Meteor., 9, 143-148.

Winkler, R. L. and A. H. Murphy, 1976: Point and area precipitation probability forecasts: some experimental results. Mon. Wea. Rev., 104, 86-95.

Winkler, R. L., A. H. Murphy, and R. W. Katz, 1977: The consensus of subjective probability forecasts: Are two, three, . . . heads better than one? Preprints, Fifth Conference on Probability and Statisics in Atmospheric Sciences, Las Vegas, American Meteological Society, Boston, 57-62.

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. Mon. Wea. Rev., 104, 1209-1214.

TERMS EXPLAINED

1. **Probability.** The chance that a prescribed event will occur, represented as a number ranging from 0 to 1. The probability of an impossible event is 0.0, that of an inevitable event is 1.0. The percentage equivalent (0 to 100%) is frequently substituted when discussing probabilities; however, the decimal equivalent (0 to 1) should be used when performing mathematical computations.

2. **Climatological Probability.** The probability that an event will occur based on extensive historical observations or experimental data. The simplest form of climatological probability (commonly called climatic frequency) is the number of occurrences of an event divided by the sum of the number of occurrences and non-occurrences over a given time period. More complex forms of climatological probability frequently use climatic models when historical observations are not available. In these cases, the models are used to obtain estimated climatological probabilities of the desired event.

3. **Sample Climatological Probability.** The climatological probability based on observations that are made only during a sample period. Examples are climatological probabilities based on one month's data.

4. **Objective Probability.** The probability that an event will occur based on a fixed set of rules which produce a unique and reproducable outcome. The rules may be derived by empirical or theoretical considerations or a combination of both.

5. **Subjective Probability.** A personal estimate of the probability that an event will occur. Subjective probability estimates give good results, if the individual knows the forecast problem (dynamics of the situation, climatology of the event, etc.) and is aware of basic probability laws and limitations of forecast skill. Subjective probability forecasts may not be reproducible.

6. **Event.** A specific occurrence that is defined by a weather element(s), time, location, and/or duration; e.g., visibility less than one mile in the period 1700-2000Z lasting more than 30 minutes at Scott AFB. Some events do not require all of the above specifications; e.g., rain at Offutt AFB at 0600Z.

7. **Probability Forecast.** Meteorological advice consisting of two parts—a well defined weather event and the expectation that the event will occur.

8. **Post Agreement.** A measure of how often an event occurs when it was forecast (forecast hits divided by total forecasts). This is a measure of categorical forecasting reliability.

9. **Prefigurance.** A measure of how often an event was forecast when it occurred (forecast hits divided by total occurrences). This is a measure of categorical forecasting capability.

10. **Correlation.** The measure of how well the forecasts agree with the observed weather. Correlation values range from -1 to +1, where -1 is perfect negative correlation, 0 is no correlation, and +1 is perfect positive correlation. (Reference AWS TR 75-259).

11. **Sharpness.** The degree of certainty of a probability forecast. A set of forecasts containing only 0% and 100% probability values has perfect sharpness. Zero sharpness occurs if all forecasts are for a probability value equal to the sample climatology.

12. **Reliability.** The degree to which forecast probabilities resemble the observed frequency for each forecast probability value or interval. For example, an event would occur 80% of the time for a series of perfectly reliable 80% probability forecasts.

13. **Decision Theory.** A set of rules designed to use probabilities and other information to make an optimal decision: information about the state of nature (a weather forecast), and information (utility, value, expense, regret, etc) on the outcome (consequence) of the decision. This information is usually given in the form of a utility matrix.

14. **Utility.** The value a decision maker associates with a given outcome with respect to other possible outcomes. It may be based on monetary value alone, or other factors which influence the decision maker's order of preference for the outcomes.

15. **Utility Matrix.** (Also called decision matrix, cost-loss matrix, expense matrix, payoff matrix, value matrix, etc, depending upon the writer and the way outcomes are quantified). A two dimensional array arranged in rows and columns. Normally, rows represent possible courses of actions (strategies, options, decisions) and columns represent the different states of nature (weather categories or thresholds). Entries at intersections of each row and column represent the outcome (utility, cost, loss, expense, payoff, value, regret, or opportunity) associated with each course of action and state of nature pair.

16. **Critical, Threshold, or Breakeven Probability.** The probability above which it is cost or mission effective for a decision maker to take a specific action, i.e., the long-term positive utility (value, payoff, etc.) is maximized and the negative utility (cost, loss, expense, regret, etc.) is minimized. Critical probability serves as the threshold which, when exceeded, generates a decision to act. It may be based on monetary value or other measures of utility. When weather is the only factor affecting the decision, the critical probability must be stated in terms of the weather event which will cause action to be taken, e.g., hangar aircraft when the probability of hail exceeds a critical probability of 10%. When other variable, non-weather mission factors affect the decision, the customer may use a critical probability stated in terms of mission success.

17. Mission Success Indicator (MSI). The probability that a mission will succeed. An MSI is tailored to a specific decision. It includes both weather (probability forecasts) and non-weather elements that are needed to make an optimal decision.

18. Weather Impact Indicator (WII). A WII is the weather input for decision assistance. It is the probability of exceeding a particular threshold of a given weather event or the probability distribution of the weather event. Customers can combine the WII with non-weather parameters to calculate a Mission Success Indicator (MSI) for use in decision making.

19. Climatological Weather Impact Indicator (CWII). A WII based on climatological probabilities rather than forecasts. CWIIs are useful for planning military operations, such as scheduling events or selecting areas or routes.

20. Simulated Weather Impact Indicator (SWII). An SWII is produced by using a model which simulates the variability of observed and forecast weather for specified climatic regimes. SWIIs can be used independently (or combined with non-weather factors to produce simulated MSIs) to study the impact of weather and weather forecasts on operations, for training aids and illustrative purposes, or to assist decision makers in the optimal use of WIIs, such as determining critical probability.

SELECTING PROBABILITY INTERVALS

Any probability value from 0 to 100% can be used for forecasting purposes, but evaluation requirements make it more desirable to use a standard set of values or intervals. In addition, use of all integral values between 0 to 100% implies a precision which does not exist in subjective probability forecasting. Table A3-1 lists the standard probabilities and ranges used by NWS forecasters in both forecasts and evaluations. Table A3-2 contains a translation of the permissible values into the verbal equivalents given to the public. The criteria used by NWS in choosing these standard values were based on verification constraints, climatology of the forecast event, and the precision of forecasting skill.

Table A3-1. NWS Permissible Probability Values (NWS Operations Manual, Chapter C-91).

VALUE (%)	PROBABILITY RANGE (%)	VALUE (%)	PROBABILITY RANGE (%)
0	$P < 2$	50	$45 \leq P < 55$
2	$2 \leq P < 5$	60	$55 \leq P < 65$
5	$5 \leq P < 8$	70	$65 \leq P < 75$
10	$8 \leq P < 15$	80	$75 \leq P < 85$
20	$15 \leq P < 25$	90	$85 \leq P < 95$
30	$25 \leq P < 35$	100	$P \geq 95$
40	$35 \leq P < 45$		

Table A3-2. Verbal Equivalents of Permissible Probability Values (NWS Operations Manual, Chapter C-91).

VERBAL TERMS	EQUIVALENT VALUES
Slight or Small Chance	$P < 30\%$
Chance	30, 40, or 50%
Likely	60 or 70%
Unqualified	$P > 70\%$

a. Verification *Constraints*. *If* every possible probability value were verified individually, the task would be exceedingly tedious, and the results difficult to interpret. The latter would occur because of the few times each probability value would be used in normal sample periods (Hughes, 1965). Therefore, it is desirable to group the probabilities into intervals which correspond as close as possible to the forecast values that will be issued. It is not possible to use standard values (such as those above) all the time in forecasts involving more than two categories. Since the sum of the probabilities for the categories must equal one, when a 2% or 5% value is used, another category must make up the difference. This difficulty does not compromise the evaluation.

b. Climatological Considerations. The range of reliable probabilities should converge to the climatological frequency of the event as lead time increases. A significant imbalance of probability intervals on either side of climatology creates a psychological problem and, if too great, may force over (under) forecasting. For this reason, probability values of 2% and 5% are included in the set used by NWS (Hughes, 1965). Therefore, forecast intervals for events that occur infrequently should have choices for the forecaster on both sides of climatology.

c. Customer Precision Requirements. The interval precision need not be any more detailed than required by the customer, but it must not be any more precise than justified by forecasting skill. Forecasters generally cannot differentiate much finer than in 10% probability intervals, except for values near the extremes (Hughes, 1965).

EXPLANATION OF MATHEMATICAL SYMBOLOGY IN THE BRIER SCORE EQUATION

Mathematically, the B ier Score (PS) is expressed by the following equation:

$$PS = \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{N} (R_{ij} - D_{ij})^2$$

a. Definition of Variables

(1) PS is the Brier Score and ranges from a value of 0.0 (perfect) to 2.0 (the worst possible).

(2) N is the total number of forecasts in the set being evaluated. A forecast with any number of categories is counted as a single forecast.

(3) K is the total number of categories in each forecast (two or more). For example, a probability of rain forecast is actually one of two categories in the forecast; the other category, the probability of no rain, is implied. If probability forecasts were issued for the combined ceiling and visibility categories (A, B, C, and D) of the AWS TAF, K would be equal to four.

(4) j is the category designator used to identify the category to which the values of R_{ij} and D_{ij} belong when the equation is expanded. It takes on all integral values from one to K.

(5) i designates the numerical order in which the forecasts will be evaluated. It ranges from one (the first) to N (the last forecast in the set).

(6) R_{ij} is the probability value assigned to category j of the i^{th} forecast. For example, $R_{42} = .9$ means that the probability for category 2 of the fourth forecast is 90%; $R_{21,3} = .3$ means that the probability for category 3 of the 21st forecast is 30%; etc. According to a law of probabilities, the sum of the probabilities for each category in a forecast must equal one; e.g., if there are only two categories involved and $R_{21} = 0.9$, then R_{22} must equal 0.1.

(7) D_{ij} is the "observed" probability and equals 1.0 if category j occurred for the i^{th} forecast; otherwise, D_{ij} equals zero. In a single forecast, only one category will have a value of $D_{ij} = 1$, and it will be the category in which the event occurred. D_{ij} in all other categories of that forecast will equal zero, regardless of the number of categories it contains.

b. Explanations

(1) $\sum_{i=1}^{N}.$ This is the symbol for the capital Greek letter "sigma." It means to sum or add the expressions that would follow for all values included in the index or subscript, i, which varies from 1 to N. Assuming N = 4,

$$\sum_{i=1}^{N=4} i = 1 + 2 + 3 + 4 = 10 \tag{A4-2}$$

$$\sum_{i=1}^{4} R_i = R_1 + R_2 + R_3 + R_4$$

$$\sum_{i=1}^{4} D_i = D_1 + D_2 + D_3 + D_4$$

$$\sum_{i=1}^{4} (R_i - D_i)^2 = (R_1 - D_1)^2 + (R_2 - D_2)^2 + (R_3 - D_3)^2 + (R_4 - D_4)^2$$

(2) $\sum_{j=1}^{K}.$ Explanation is similar to that above.

(3) $\frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{N}$. This means that two summations must be made and the constant, 1/N, multiplied by the result. The normal procedure is to set j=1 in the first sigma, and then sum all cases for i=1 to N using the second sigma. The procedure is then repeated each time for j=2, j=3 through the value j=K. Assuming N=4 and K=4,

$$\frac{1}{N} \sum_{j=1}^{K=4} \sum_{i=1}^{N=4} (R_{ij}-D_{ij})^2 = \frac{1}{4} \{[(R_{11}-D_{11})^2 + (R_{21}-D_{21})^2 + (R_{31}-D_{31})^2 + (R_{41}-D_{41})^2]$$
$$+[(R_{12}-D_{12})^2 + (R_{22}-D_{22})^2 + (R_{32}-D_{32})^2 + (R_{42}-D_{42})^2]$$
$$+[(R_{13}-D_{13})^2 + (R_{23}-D_{23})^2 + (R_{33}-D_{33})^2 + (R_{43}-D_{43})^2]$$
$$+[(R_{14}-D_{14})^2 + (R_{24}-D_{24})^2 + (R_{34}-D_{34})^2 + (R_{44}-D_{44})^2]\} \quad \text{(A4-3)}$$

Note that values in the first set of brackets represent the contribution to the total from category 1; the second bracket, the contribution from category 2; etc.

c. Example, the last equation above represents the expanded form of the Brier Score equation for calculating scores for four categories of a set of four forecasts. Sample values for four such forecasts are given in Table A2-1 below. They will be substituted in the equation to illustrate computational procedures.

Table A4-1. Verification for Four Forecasts.

FCST # (i)	OBSVD CAT	CATEGORY 1		CATEGORY 2		CATEGORY 3		CATEGORY 4	
		FCST PROB (R_{i1})	OBSVD PROB (D_{i1})	FCST PROB (R_{i2})	OBSVD PROB (D_{i2})	FCST PROB (R_{i3})	OBSVD PROB (D_{i3})	FCST PROB (R_{i4})	OBSVD PROB (D_{i4})
1	4	.0	0	.0	0	.1	0	.9	1
2	4	.0	0	.0	0	.0	0	1.0	1
3	4	.0	0	.0	0	.0	0	1.0	1
4	4	.0	0	.1	0	.2	0	.7	1

Substituting values for R_{ij} and D_{ij}:

$$PS = \frac{1}{4} \sum_{j=1}^{4} \sum_{i=1}^{4} (R_{ij}-D_{ij})^2 = \frac{1}{4}\{[(.0-0)^2 + (.0-0)^2 + (.0-0)^2 + (.0-0)^2] \qquad (A4-4)$$

$$+[(.0-0)^2 + (.0-0)^2 + (.0-0)^2 + (.1-0)^2]$$

$$+[(.1-0)^2 + (.0-0)^2 + (.0-0)^2 + (.2-0)^2]$$

$$+[(.9-1)^2 + (1.0-1)^2 + (1.0-1)^2 + (.7-1)^2]\}$$

$$PS = \frac{1}{4} \sum_{j=1}^{4} \sum_{i=1}^{4} (R_{ij}-D_{ij})^2 = \frac{1}{4}\{[0] + [.01] + [.05] + [.1]\}$$

$$= 0 + .003 + .013 + .025$$

$$PS = .041$$

In the next to last line above, each of the four values represents the Brier Scores for the respective category, K = 1, 2, 3, 4. Summing these indivudual scores gives the total Brier Score. Refer back to the point where we substituted values into the equation above. The word meaning of those mathematical symbols is simply this: the Brier Score is the average of the squares of the forecast errors. It is an average, because we divide by the number of forecasts involved, and the values we average are the squares of the differences between the forecast and observed probabilities.

d. Alternate Methods. By now it should be obvious that calculation of the Brier Score is very unwieldy using the above method when a large number of forecasts are involved. Tables 3-3, 3-4 and 3-5 and related discussions in the main text explain how the procedures can be greatly simplified using tabular formats to perform the computations. The data in Table A4-1 above are the same as the first four forecasts used in Table 3-4 of the text; therefore, the two methods may be compared directly.

TABLE OF PARTIAL BRIER SCORES

A basic understanding of the mathematical meaning of the Brier Score equation is necessary regardless of how one actually computes the score. However, there are several shortcuts that can be devised to simplify the computations. Some of those were described in the basic part of the pamphlet. Table A5-2 is one example. Specifically, it eliminates the need to repeatedly compute specific values of $(R_{ij} - D_{ij})^2$, the penalty points associated with the Brier Score, and allows the data to be put in tabular format for easy computation. The example verification summary given in Table A5-1 below is used to illustrate procedures for extracting partial Brier Scores (PS_p) from Table A5-2

Table A5-1. Brier Score Computation Using Table of Partial Brier Scores.

FCST PROB (R_{ij})	TOTAL # OF FCSTS	OCCURRENCES (D_{ij} = 1)		NONOCCURRENCES (D_{ij} = 0)		ΣPS_p
		#FCSTS (n)	PS_p	# FCSTS (n)	PS_p	$[n(R_{ij}-D_{ij})^2]$
1.0	7	5	0.00	2	2.00	2.00
.8	4	2	0.08	2	1.28	1.36
.6	4	1	0.16	3	1.08	1.24
.4	1	1	0.36	0	0.00	0.36
.2	6	1	0.64	5	0.20	0.84
.0	9	0	0.00	9	0.00	0.00
TOTAL	31	10	1.24	21	4.56	5.80

$$PS = \frac{2}{N} \sum_{i=1}^{K} (R_i - D_i)^2 = \frac{2}{31} (5.8) = .374$$

a. Instructions. Table A5-2 gives values for $n(R_{ij} - D_{ij})^2$, where n is the number of forecasts in the probability interval corresponding to the value of R_{ij} for either occurrences (D_{ij} = 1) or nonoccurrences (D_{ij} = 0) of the event.

(1) To determine penalties (PS_p) for event occurrences, use forecast probabilities (R_{ij}) in the top column heading (D_{ij} = 1). Locate the appropriate value for R_{ij}, then go down the column to the row corresponding to the number of forecasts (n) in which the event occurred. In Table A3-1 there were five forecasts with a probability of 1.0. The penalty is 0.00. Two forecasts for a probability of .8 give a penalty of 0.08, etc, for all other occurrences.

(2) Penalties for nonoccurrences of the event use forecast probabilities (R_{ij}) in the second column heading (D_{ij} = 0). Other procedures for extracting the penalties (PS_p) are the same as above.

(3) Sum the Partial Brier Score obtained in both steps above and divide by the total number of forecasts issued (N) to obtain the Brier Score for that one category. If the forecast is for a two category system, multiply the result by 2 to obtain the total Brier Score (reference para 3-6). For three or more categories, determine Brier Scores for each category as above (do not multiply by 2) and sum them to obtain the total Brier Score (reference Table 3-5 for an example).

b. One is not restricted to using only the probability values given in the tables. Other intermediate values could be added, if needed. Further, there is nothing magic about where the tables stopped with values of (n). Expand the table if you routinely need partial scores for a larger number of forecasts (n).

TABLE A5-2. PARTIAL BRIER SCORES $\left[PS_p = n\,(R_{ij} - D_{ij})^2\right]$

	FORECAST PROBABILITY												
OCCURRENCE (D_{ij} = 1)	1.00	.95	.9	.8	.7	.6	.5	.4	.3	.2	.1	.05	.0
NON OCCURRENCE (D_{ij} = 0)	.0	.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	.95	1.00
n (# OF FORECASTS IN EACH PROBABILITY INTERVAL)													
1	.00	.00	.01	.04	.09	.16	.25	.36	.49	.64	.81	.90	1.00
2	.00	.01	.02	.08	.18	.32	.50	.72	.98	1.28	1.62	1.81	2.00
3	.00	.01	.03	.12	.27	.48	.75	1.08	1.47	1.92	2.43	2.71	3.00
4	.00	.01	.04	.16	.36	.64	1.00	1.44	1.96	2.56	3.24	3.61	4.00
5	.00	.01	.05	.20	.45	.80	1.25	1.80	2.45	3.20	4.05	4.51	5.00
6	.00	.02	.06	.24	.54	.96	1.50	2.16	2.94	3.84	4.86	5.42	6.00
7	.00	.02	.07	.28	.63	1.12	1.75	2.52	3.43	4.48	5.67	6.32	7.00
8	.00	.02	.08	.32	.72	1.28	2.00	2.88	3.92	5.12	6.48	7.22	8.00
9	.00	.02	.09	.36	.81	1.44	2.25	3.24	4.41	5.76	7.29	8.12	9.00
10	.00	.03	.10	.40	.90	1.60	2.50	3.60	4.90	6.40	8.10	9.03	10.00
11	.00	.03	.11	.44	.99	1.76	2.75	3.96	5.39	7.04	8.91	9.93	11.00
12	.00	.03	.12	.48	1.08	1.92	3.00	4.32	5.88	7.68	9.72	10.83	12.00
13	.00	.03	.13	.52	1.17	2.08	3.25	4.68	6.37	8.32	10.53	11.73	13.00
14	.00	.04	.14	.56	1.26	2.24	3.50	5.04	6.86	8.96	11.34	12.64	14.00
15	.00	.04	.15	.60	1.35	2.40	3.75	5.40	7.35	9.60	12.15	13.54	15.00
16	.00	.04	.16	.64	1.44	2.56	4.00	5.76	7.84	10.24	12.96	14.44	16.00
17	.00	.04	.17	.68	1.53	2.72	4.25	6.12	8.33	10.88	13.77	15.34	17.00
18	.00	.05	.18	.72	1.62	2.88	4.50	6.48	8.82	11.52	14.58	16.25	18.00
19	.00	.05	.19	.76	1.71	3.04	4.75	6.84	9.31	12.16	15.39	17.15	19.00
20	.00	.05	.20	.80	1.80	3.20	5.00	7.20	9.80	12.80	16.20	18.05	20.00
21	.00	.05	.21	.84	1.89	3.36	5.25	7.56	10.29	13.44	17.01	18.95	21.00
22	.00	.06	.22	.88	1.98	3.52	5.50	7.92	10.78	14.08	17.82	19.86	22.00
23	.00	.06	.23	.92	2.07	3.68	5.75	8.28	11.27	14.72	18.63	20.76	23.00
24	.00	.06	.24	.96	2.16	3.84	6.00	8.64	11.76	15.36	19.44	21.66	24.00
25	.00	.06	.25	1.00	2.25	4.00	6.25	9.00	12.25	16.00	20.25	22.56	25.00
26	.00	.07	.26	1.04	2.34	4.16	6.50	9.36	12.74	16.64	21.06	23.47	26.00
27	.00	.07	.27	1.08	2.43	4.32	6.75	9.72	13.23	17.28	21.87	24.37	27.00
28	.00	.07	.28	1.12	2.52	4.48	7.00	10.08	13.72	17.92	22.68	25.27	28.00
29	.00	.07	.29	1.16	2.61	4.64	7.25	10.44	14.21	18.56	23.49	26.17	29.00
30	.00	.08	.30	1.20	2.70	4.80	7.50	10.80	14.70	19.20	24.30	27.08	30.00
31	.00	.08	.31	1.24	2.79	4.96	7.75	11.16	15.19	19.84	25.11	27.98	31.00
32	.00	.08	.32	1.28	2.88	5.12	8.00	11.52	15.68	20.48	25.92	28.88	32.00
33	.00	.08	.33	1.32	2.97	5.28	8.25	11.88	16.17	21.12	26.73	29.78	33.00

DETERMINING UTILITIES IN TERMS OF REGRET

1. INTRODUCTION. The concept of utilities is rather simple to understand, but procedures for actually determining utility values can be difficult to grasp. The following extract from Selvidge's Technical Report 76-12, Rapid Screening of Decision Options (1976) vividly illustrates how one might go about developing utilities in terms of regret. Although the example used is not a meteorological application, the principles involved in a meteorological decision problem are the same.

a. **Warsaw Pact Attack Example.** The first step for rapidly evaluating decision options is to describe the decision problem in a simplified format. The following example provides a concrete application of this format. The problem analyzed is one which might be faced by a NATO decision maker.

Suppose that intelligence information indicates that there is a build-up of Warsaw Pact forces in Eastern Europe and the Western USSR. The uncertain event of interest is whether or not these forces will invade NATO countries. The decision to be made is: What alert posture should NATO assume? The decision about the extent of the alert must be made before the intentions of the Warsaw Pact forces are known for certain. If the NATO commander is considering four alternative levels of alert: Maintain status quo, military vigilance, simple alert, and reinforced alert, then the decision problem can be structured in the simplified decision-tree format shown in Figure A6-1. (For additional information on decision trees, see Attachment 9).

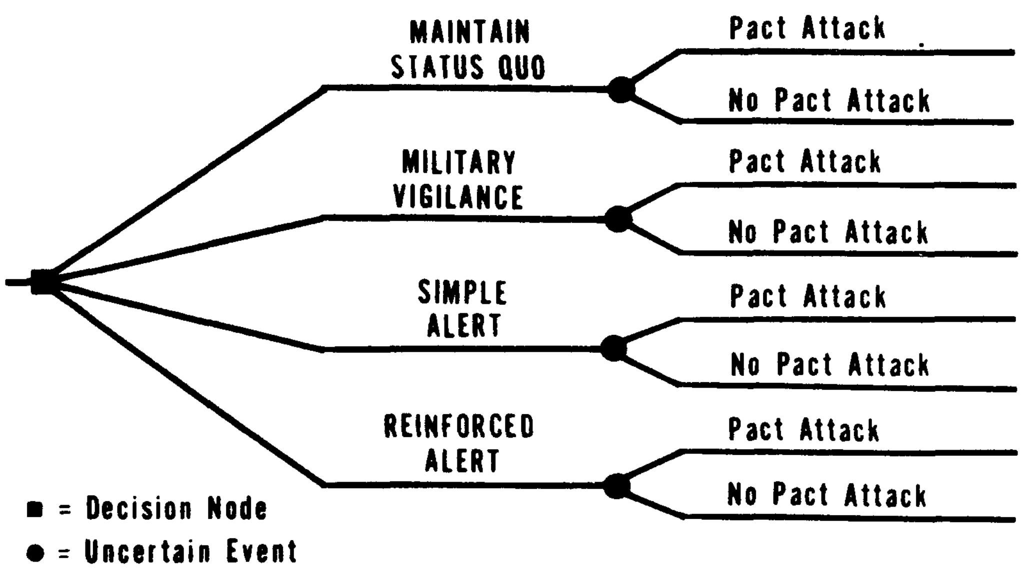


Figure A6-1. Warsaw Pact Attack Example--Simplified Format.

In matrix form, this decision problem has the rows and columns shown in Table A6-1.

VALUE MATRIX

DECISION OPTIONS	UNCERTAIN EVENT: IS AN ATTACK PLANNED ?	
	OUTCOMES	
	PACT ATTACK	NO PACT ATTACK
MAINTAIN STATUS QUO		
MILITARY VIGILANCE		
SIMPLE ALERT		
REINFORCED ALERT		

PROBABILITIES

PACT ATTACK	NO PACT ATTACK

Table A6-1. Warsaw Pact Attack Example in Matrix Form.

The uncertain event has been defined as whether or not the Warsaw Pact forces are planning to attack. The simplifying assumption is made that the intention of the Warsaw Pact forces does <u>not</u> depend on whether the NATO forces maintain or increase their level of alert. Therefore, the probabilities of attack and no attack need to be estimated only once and will not change after the NATO decision maker selects among the decision options.

b. Value Structure. The final step in structuring a decision problem is to identify the important factors that describe the possible consequences of outcomes and options and to determine how happy or unhappy the decision maker expects to be with a particular decision. These factors are the dimensions on which the decision maker's satisfaction with different combinations of options and outcomes is measured. For some problems, a great many descriptors can be applied to the consequences. In that case, the analyst should restrict consideration to the factors of primary importance. By definition, there cannot be too many of these. Often fewer than a half dozen factors are sufficient to describe the consequences. The consequences of many business decisions, for example, can be described simply in monetary terms. For social and military decision making, however, factors such is "political implications" or "lives lost" may be important. Besides selecting these important factors from among the many possible, the decision maker must also assign an "importance weight" to each factor. These weights indicate relative importance among the different factors and are used to combine ratings on each of the different dimensions into a single summary measure of the value. In the Warsaw Pact attack example, a military operations expert described in some detail what activities would be entailed in each of the options (maintaining the status quo through reinforced vigilance) and what the probable consequences of these activities would be both for the case of a Pact attack and for no Pact attack. After listing these consequences, the military expert concluded that they could be grouped into three general categories:

- o **Alert Cost** (e.g., cost of deploying additional forces, assuming control of civilian transportation);
- o **Political Cost** (e.g., embarrassment of being wrong if NATO forces prepare for an attack which never materializes); and
- o **Military Risk** (e.g., expected military loss—lives, equipment, territory, etc.—if the attack occurs and NATO is unprepared).

These categories become the value dimensions of interest. To fill the value matrices, three basic matrices are set up, each representing one of the value dimensions. Each option and event outcome combination is rated on each of these dimensions. Then a fourth matrix, the "combination valuation," which is the weighted sum of the measures in each of three categories, is formed.

2. ASSESSING INPUTS. The analyst or the user must quantify the uncertainty about the event outcomes in terms of probabilities and must also express the desirability (or, alternatively, the lack of satisfaction) of the option and outcome combinations on the dimensions identified earlier. Because the outcomes are defined so that they are independent of the options (i.e., do not change as a function of the option), the probability assessment may take a relatively small proportion of the effort devoted to preparing inputs. The value assessments are generally much more difficult and time-consuming. Initially, however, both of these inputs can be approximations rather than the most accurate possible reflections of uncertainty and value.

a. **Probabilities of the Outcomes.** Among statisticians and others interested in the study and use of probabilities as a measure of uncertainty, there are presently two main schools of thought about how probabilities should be defined. One is the "objectivist" or "frequentist" school which maintains that the probabilities of outcomes can only be found from the long-run relative frequency of occurrence of outcomes of identical events. The other is the "subjectivist" or "personalist" school which says that probability is a measure of someone's degree of belief that an outcome will occur. The latter definition is generally used by decision analysts since rarely is the decision problem studied one which has occurred exactly in the same form many times in the past. For instance, in the Warsaw Pact attack example, we cannot look at the past and say that identical circumstances have occurred repeatedly and that sometimes the Pact attacked and sometimes it did not. Rather than trying to get a relative frequency measure of the probability, the analyst or user of the procedure tries to quantify the degree of belief of some expert. Many experiments have been carried out in order to arrive at guidelines for ways of eliciting this probabalistic information in different circumstances. The expert, or a group of experts, is asked questions like: "Which outcome is most likely?" and "How many times more likely is this than the next most likely?" "Than the least likely?" Eventually the replies can be consolidated into probabilities (or percentages) for the different outcomes.

In the Warsaw Pact attack example, the expert considered many intelligence reports of recent Soviet domestic affairs, Soviet activities in the Mediterranean, Warsaw Pact countries' military maneuvers, and the like. Considering this information, the expert eventually arrived at probabilities of 0.10 for the outcome Warsaw Pact attack and 0.90 for the outcome no attack. (The list of outcomes whose probabilities are assessed must be exhaustive; that is, their probabilities must add to 1.00 or to 100 when expressed as a percentage.)

The assessment of the probabilities is more complicated if:

- o The assessor is periodically receiving new information and would like to update the probabilities to reflect this information; or
- o The uncertain event of interest is actually the last of a series of other uncertain events and its probabilities are conditioned by how the other events turn out.

b. **Values of the Option-Outcome Combinations.** The structure of the decision problem determines the value dimensions and the option-outcome combinations for whose consequences the values must be assessed. For the Warsaw Pact attack example, the user provides the numbers to fill the matrices displayed in Table A6-2. Assessing these values can be difficult because, in most simplified examples such as this, each value *dimension* is a composite and so may have no natural scale. When this is the case, an arbitrary scale is established. The user or expert whose judgments are to be quantified is then asked a series of questions which require considerable thought to answer. These questions are designed to elicit

ALERT COST

	PACT ATTACK	NO PACT ATTACK
MAINTAIN STATUS QUO		
MILITARY VIGILANCE		
SIMPLE ALERT		
REINFORCED ALERT		

POLITICAL COST

	PACT ATTACK	NO PACT ATTACK
MAINTAIN STATUS QUO		
MILITARY VIGILANCE		
SIMPLE ALERT		
REINFORCED ALERT		

MILITARY RISK

	PACT ATTACK	NO PACT ATTACK
MAINTAIN STATUS QUO		
MILITARY VIGILANCE		
SIMPLE ALERT		
REINFORCED ALERT		

Table A6-2. Value Matrices for the Warsaw Pact Attack Example.

the user's feelings about how the option-outcome combinations rate on the selected arbitrary scale.

There are two general types of arbitray scales, either of which can be used in a decision problem. One is an absolute scale called "payoffs," the other a relative scale called "regret." The payoff scale is described briefly. The regret scale, which is the recommended scale for the decision problems discussed here, is described at length.

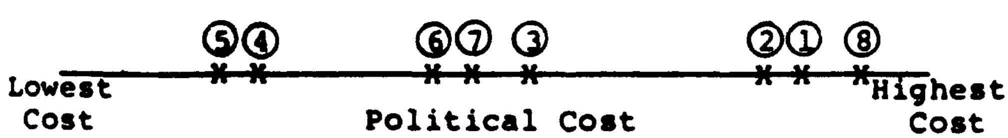
(1) **Payoffs** - Consider the value dimension "political cost" in the Warsaw Pact attack example. There are eight possible option-outcome combinations shown in Table A6-3.

	PACT ATTACK	NO PACT ATTACK
MAINTAIN STATUS QUO	1	5
MILITARY VIGILANCE	2	6
SIMPLE ALERT	3	7
REINFORCED ALERT	4	8

Table A6-3. Possible Option Outcome Combinations.

They range from maintaining the status quo and a pact attack (combination 1) to reinforced alert and no attack (combination 8). One way to think of the value problem is by imagining an arbitrary political cost scale along which the assessor must scatter points representing option-outcome combinations in positions that show their relative

desirability. A hypothetical scattering is shown below. The circled numbers represent the option-outcome combinations (the cells) of Table A6-3.



The values read off this scale fill the "political cost" value matrix, which is analogous to the payoff table prepared in an elementary decision analysis exercise.

(2) **Regret** - An alternative way of expressing the "political cost" value is to consider one column of the matrix at a time (corresponding to one of the outcomes of the uncertain event) and within the column to make judgments about the relative cost (value) of different possible options under that outcome as compared to the best option. For instance, if, on the assumption that there will be a pact attack, what is the best option? And then what are the values of all the other options compared to that best option? This process is analogous to preparing a "regret table" in elementary decision analysis. Many users find it easier to think about "regrets" under a specific assumption about the outcome than to make judgments about payoffs where the users must consider both different outcomes and different options at once. For this reason the regret scale is used in these examples.

In order to respond to questions like "How"—in terms of regret—"does the value of option 1 compare to that of option 2?" The units in which regret is measured must be chosen. The decision is an arbitrary one which can be handled as shown in steps 1 and 2 below. Suppose that you are the assessor whose values are being elicited.

(3) **Rules for filling a regret matrix**

Arbitrary Establishment of the Units

1. If you make the decision which is best for a particular event outcome, then you have no regret. Therefore, within each column, identify the option that would be optimal if the outcome of the uncertain event were that indicated by the column under consideration. Set the regret of that cell equal to zero. When you have finished, each column will contain a cell with zero in it. This cell establishes one end of the regret scale within each column.
2. Within each column, identify the worst option. Then for each column think about how you feel on the dimension of value being considered about going from the best to the worst option in that column. You have no regret with the best option, but you may have a great deal of regret with the worst option. Then, for each column, decide which of these transitions from best to worst option involves the greatest incremental increase in regret. Assign a value of -100 to the worst option cell in the column where this increase in regret is greatest.

Relative Value Assessments

3. For the column containing both a zero value and a -100 value, assign values between 0 and -100 to the rest of the cells which reflect the relative regret of each cell compared to the -100 cell. For example, if the amount of your regret in going from the zero cell to another cell is about 1/4 of that for going from the zero cell to the -100 cell, then the other cell should have a value of about -25.
4. Next, consider the minimum-to-maximum regret cells of another column in comparison to the 0 to -100 cells of the previous column. Use your feelings about this regret difference to establish the value of the maximum regret cell for the new column. For example, if you think it is about half as bad to go from the best to the worst option for this new column as going from the 0 to the -100 options in the previous column, then the maximum regret cell of the new column should have a value of -50.

Repeated Assessments

5. The procedure of step 3 is repeated to get all the cell values within each column, and that of step 4 is repeated to determine the worst regret value in each column before the intermediary cells are filled.

Adjustments

6. Once the cells have been filled, various pair-wise comparisons are made to test and increase the consistency of the assessments. In these pair-wise comparisons, the difference between the regret values for one pair is compared to the difference between values for another pair. These comparisons can be made both within a column and across columns.

(4) **Regret assessment example: Warsaw Pact attack** - The regret assessment is more easily understood in the context of a particular example than by merely listing the assessment rules. Furthermore, specific problems sometimes have special features that reduce the number of judgments needed. Consider first the political cost dimension of the Warsaw Pact attack problem. The regret matrix to be filled is shown in Table A6-3A. Following the rules explained above, the assessor looks first at each column separately to find its zero-regret cell. If the event outcome is attack, the best option from a political standpoint is the maximum alert posture, namely reinforced alert; this option avoids, for example, the loss of face in being taken by surprise; consequently, this cell is given a zero value. On the assumption there will be no attack, on the other hand, the best option (zero regret) from the standpoint of political costs is simply to maintain the status quo. Table A6-4B shows the appearance of the regret matrix after these judgments are made.

Next, the worst option (maximum regret) under each outcome is noted. In this case the worst decisions are if there is an attack, maintain status quo, and if there is no attack, reinforced alert. The assessor must decide whether there is more regret, on the political cost dimension, in going from the best to worst option under the attack outcome compared to going from the best to worst option under the no attack ou... ..e. Another way of phrasing this question is, "Is it a bigger mistake politically to have failed to go on alert if there is an attack or to have gone on alert when there is no attack?" It happens in this particular example, that because the assessor feels that those two shifts are equally bad, the "worst option" cell in each column is assigned a value of -100. (See Table A6-4C).

The next step in the regret assessment is to fill in the intermediary values in the column containing both a 0 and a -100. (In this example either column satisfies that requirement.) We begin with the attack column. After some thought, the assessor comes up with the values shown in Table A6-4D. These values imply that, of the total (political cost) regret possible from being wrong on the decision (i.e., selected the wrong option) the event outcome is attack, only about a tenth of that is incurred by going on a simple alert instead of a reinforced alert and about a third is incurred in choosing the military vigilance option rather than reinforced alert. One way to explain these values is that the assessor feels that the political cost of being in a less than maximum alert

POLITICAL COST

A.

	ATTACK	NO ATTACK
MAINTAIN STATUS QUO		
MILITARY VIGILANCE		
SIMPLE ALERT		
REINFORCED ALERT		

B.

	ATTACK	NO ATTACK
MAINTAIN STATUS QUO		0
MILITARY VIGILANCE		
SIMPLE ALERT		
REINFORCED ALERT	0	

C.

	ATTACK	NO ATTACK
MAINTAIN STATUS QUO	−100	0
MILITARY VIGILANCE		
SIMPLE ALERT		
REINFORCED ALERT	0	−100

D.

	ATTACK	NO ATTACK
MAINTAIN STATUS QUO	−100	0
MILITARY VIGILANCE	−35	
SIMPLE ALERT	−10	
REINFORCED ALERT	0	−100

E.

	ATTACK	NO ATTACK
MAINTAIN STATUS QUO	−100	0
MILITARY VIGILANCE	−35	−20
SIMPLE ALERT	−10	−50
REINFORCED ALERT	0	−100

Table A6-4. Application of Rules for Filling Regret Matrix.

posture (compared to being in the maximum alert posture) if an attack develops is quite a bit less serious than having maintained the status quo. In other words, having gone to some level of greater alert (exactly which level is not as crucial) is much better from a political viewpoint than having done nothing.

The regret values for the second column can be assessed directly since this column also contains both 0 and -100 values. (If the regret of going from best to worst decision here had been less than that of going from best to worst in column 1, then the maximum regret value for this column would have been assessed as something less than -100, for example, -67 if the maximum amount of regret in column 2 were thought to be about 2/3 that of column 1. The rest of the values in column 2 are estimated following the same procedure as described for column 1. The results of this assessment appear in Table A6-4E.

All the assessments in the matrix can then be checked (and adjusted if necessary) by making pair-wise assessments of the values. For example, the assessor's feelings of regret should be twice as serious going from the 0 to -20 cells in column 2 as going from the 0 to -10 cells in column 1.

Taken together, these feelings imply that, in the opinion of the assessor, having gone to military vigilance when in fact there is no attack is twice as serious a mistake as having gone only to simple alert when an attack does occur.

The regret matrices for the other two dimensions, military risk and alert cost, are assessed in the same manner as the political cost. In assessing military risk, for example, the zero regret cells are identified as shown in Table A6-5. Furthermore, the assessor concludes here that, **militarily speaking,** there is no regret in being over-prepared for an attack which does not materialize. For this reason all the cells in column 2 are zero. The worst option on the assumption that there is an attack is to maintain the status quo; accordingly, that cell receives a regret value of -100 (see Table A6-5B). The rest of the values were assessed as shown in Table A6-5C.

A.

	ATTACK	NO ATTACK
MAINTAIN STATUS QUO		0
MILITARY VIGILANCE		
SIMPLE ALERT		
REINFORCED ALERT	0	

B.

MAINTAIN STATUS QUO	-100	0
MILITARY VIGILANCE		0
SIMPLE ALERT		0
REINFORCED ALERT	0	0

C.

MAINTAIN STATUS QUO	-100	0
MILITARY VIGILANCE	-45	0
SIMPLE ALERT	-15	0
REINFORCED ALERT	0	0

Table A6-5. Military Risk.

The final value dimension, alert cost, is meant to be a measure of the out-of-pocket costs of going from the status quo to the various levels of alert. However, rather than trying to figure out these costs in dollars, they will also be approximated by regret on a scale of 0 to -100. If the objective is to minimize regret on a cost dimension, then the best option (that having the lowest cost) is to maintain the status quo and the worst is to go to the reinforced alert. These regrets are, therefore, 0 to -100, respectively. Since the cost and, consequently, the regret remains the same, whether there is an attack or not, both columns of the regret matrix for alert cost will be the same. The values obtained during this assessment for the different cells of the matrix are shown in Table A6-6.

	ATTACK	NO ATTACK
MAINTAIN STATUS QUO	0	0
MILITARY VIGILANCE	-30	-30
SIMPLE ALERT	-70	-70
REINFORCED ALERT	-100	-100

Table A6-6. Alert Cost.

In applying the general rules for filling a regret matrix to this example, several special features of the example became apparent. These were:

- o In the political cost matrix, the amount of regret incurred in going from the best to worst option under one outcome (attack) was felt to be the same as the amount incurred in going from the best to worst option under the other outcome (no attack). (Both columns contained a 0 and a -100.)

- o For the military cost matrix under the outcome of no attack, none of the non-optimal options resulted in any regret when compared to the optimal one. (All the entries of the second column were 0.)

- o In the alert cost matrix, the amount of regret was the same regardless of which outcome was assumed to occur. (Column 1 is identical to column 2.)

One feature of measures of regret which should be kept in mind when regret matrices are used is that making comparisons of values across columns is somewhat tricky. Regret values within a column are all measures of the value of a cell **relative** to that of the optimal cell for that column. The basis for these relative values must be kept in mind for a comparison of regret values across columns. If two cells in different columns both contain the regret value of -35 , for instance, then the assessor feels as bad about going from the optimal cell in one column to its -35 cell as about going from the optimal cell in the other column to its -35 cell. This equivalence is in contrast to the interpretation of the entries in a payoff matrix. For a payoff matrix the values in the cells are measured in absolute terms. If two cells in different columns have the same payoff, then the assessor feels equally good (or bad) about **being** in either of the states. For two regret cells having equal values, the assessor feels equally good about the **transition** to that cell from the optimal cell in its column. Statements involving the comparison of incremental regrets can also be made. For example, if the difference between two regrets in one column is, say, 20 this is the same amount of regret as that between any two regrets in another column which also differ by 20.

c. **Weights for the Value Dimensions.** After the assessor's feelings about regret have been elicited for each of the different value dimensions, these figures are combined into a single value for every decision option-event outcome combination. This composite regret matrix, called the "combined valuation," is formed by taking a weighted average of the matrices over the different value dimensions. The average is weighted because in most examples certain of the dimensions are more important than others. These weights are assessed as part of this analysis.

When values over different dimensions are expressed in terms of regret, their weights are called the "swing weights" and are estimated not by considering the overall difference in importance of one dimension compared to another, but rather by estimating the importance of a **swing** from the best (regret = 0) to worst (regret = -100) option in one column of one dimension compared to the swing from the best to worst option on another dimension. For example, consider the regret matrices in the Warsaw Pact attack case shown in Table A6-7. Now suppose the assessor first considers the military risk compared to political cost and decides that the military risk regret of going from zero (reinforced alert if an attack occurs) to -100 (maintaining status quo if an attack occurs) is twice as important as the political cost regret of going from zero to -100 under the same conditions; that is, the swing weight for military risk is twice the swing weight for political cost. Suppose further that the assessor decides that the alert cost regret from having spent the money to go from zero (maintaining status quo) to -100 (going on reinforced alert) is about equal in importance to the political cost regret of going from zero (reinforced alert if there is an attack) to -35 (military vigilance, if there is an attack). This implies that the political cost swing weight is about three times that of the alert cost. To summarize these assessments:

military risk importance = 2 x political cost importance.

political cost importance = 3 x alert cost importance.

Maintaining these relationships and normalizing the weights so that they add to 1.00 give:

Value Dimension	Importance Weight
Military risk	0.6
Political cost	0.3
Alert cost	0.1

3. **CALCULATIONS.** Once the decision problem has been structured and the inputs assessed, some straightforward calculations are made to enable the user to determine the best decision option.

a. **Combined Valuation.** By means of the importance weights discussed above, the different regret matrices are combined into a single matrix expressing the combined effects of regret on different dimensions. The result of this computation is shown in Table A6-8 on the following page. The assumptions are made that the different dimensions of value are independent and that they combine according to an additive rule. Under these assumptions, each cell in the combined valuation matrix is filled by taking the weighted average of the regret values in the corresponding cells of the three value dimension matrices. For example, the following computation produces the value of -19 in the simple alert-attack cell:

$$(-10 \times 0.30) + (-15 \times 0.60) + (-70 \times 0.10) = -19.$$

As is the case with the individual regret matrices, the values of cells in this combined matrix incorporate an understood comparison with the value of the optimal cell in each column. In the combined value matrix, however, the optimal cell for each column will not necessarily have a zero value, since the combined valuation is a weighted sum of the individual value matrices. For instance, in the Warsaw Pact attack example, the "Attack" column of the combined valuation matrix no longer has a zero entry. Before comparisons can be made of the absolute values of the regret from column to column, the zero must be restored, in this case by adding 10 to every entry in that column. (Whether this adjustment is made or not has no effect upon the choice of the optimal act since the addition of the same constant to each row of the matrix will not change which row has the smallest expected regret, the choice criterion discussed in the next section.) Without making this adjustment, however, the differences between entries within one column can be compared to entry differences in another column. For example, the amount of regret (9 units) in going from zero to -9 in the second column is the same as that of going from -10 to -19 in the first column. However, before the adjustment, the amount of regret from being in the military vigilance-attack cell (regret = -40) is not the same (-40) as that of being in the reinforced alert-no attack cell.

b. **Expected Value.** The criterion used here for indicating the best decision option is that having the smallest expected regret, measured from the values of the combined valuation matrix. Expected regret is computed for each option by multiplying the value of each outcome under that option by the outcome's probability. For example, the option "reinforced alert" has combined regret values of -10 if there is an attack and -40 if there is no attack. Weighting these values by the probabilities for the two outcomes gives:

$$(-10 \times 0.10) + (-40 \times 0.90) = -37.$$

Carrying out the computation for the other three options gives the expected regret values shown in Table A6-9. Since the smallest expected regret value is 9 (ignore the minus signs, which are included merely to remind the assessor that regret is a measure of undesirability), the associated option, maintaining the status quo, is, therefore, the optimal decision on the basis of the data input.

c. **Sensitivity.** The expected regret values for each of the four options considered here depend on the three kinds of inputs to the analysis: the regret matrices for the different dimensions, the importance weights for the different dimensions, and the probabilities. One way to obtain these input values is to spend a lot of time and effort in making the assessments. Generally, however, a more efficient way to conduct the analysis is to assess quickly some approximate numbers for use in an initial pass through the whole procedure. The final step in the option screening method then becomes a sensitivity analysis where changes are made to the inputs to see their effect upon the solution, that is, the choice of the option having the smallest expected regret.

(1) **Probabilities** - The expected regret for each option is a linear function of the corresponding row in the combined valuation matrix with the probabilities serving as coefficients. Changes in the probabilities of attack versus no attack will cause changes in the values of the expected regret and may cause a change in the optimal option, that is, which option has the smallest expected regret. Because of the linearity of the relationship, the effect of probability changes on expected regret can be easily shown graphically. In Figure A6-2, four lines are plotted. Each of these, one for each option, is an expected regret line. The points composing the line show the change in expected regret (the vertical scale) for changed values of the probability of attack (the horizontal scale). (The probability of no

	POLITICAL COST		MILITARY RISK		ALERT COST	
	ATTACK	NO ATTACK	ATTACK	NO ATTACK	ATTACK	NO ATTACK
MAINTAIN STATUS QUO	−100	0	−100	0	0	0
MILITARY VIGILANCE	−35	−20	−45	0	−30	−30
SIMPLE ALERT	−10	−50	−15	0	−70	−70
REINFORCED ALERT	0	−100	0	0	−100	−100

Table A6-7. Regret Matrices for All Three Value Dimensions.

	POLITICAL COST			MILITARY RISK			ALERT COST			COMBINED VALUATION	
IMPORTANCE WEIGHTS	0.30 ×			0.60 ×			0.10 ×				
	ATTACK	NO ATTACK		ATTACK	NO ATTACK		ATTACK	NO ATTACK		ATTACK	NO ATTACK
MAINTAIN STATUS QUO	−100	0		−100	0		0	0		−90	0
MILITARY VIGILANCE	−35	−20	+	−45	0	+	−30	−30	=	−40	−9
SIMPLE ALERT	−10	−50		−15	0		−70	−70		−19	−22
REINFORCED ALERT	0	−100		0	0		−100	−100		−10	−40

Figure A6-8. Regret Matrices for All Three Value Dimensions Combined Into a Single Matrix.

COMBINED VALUATION

	ATTACK	NO ATTACK
MAINTAIN STATUS QUO	−90	0
MILITARY VIGILANCE	−40	−9
SIMPLE ALERT	−19	−22
REINFORCED ALERT	−10	−40

X

PROBABILITY OF THE OUTCOMES

ATTACK	NO ATTACK
0.10	0.90

=

EXPECTED VALUE OF THE REGRET

−9
−12
−22
−37

* COMPUTER ROUNDS OFF FIGURES TO NEAREST INTEGER

Table A6-9. Computation of the Expected Value of the Combined Regret.

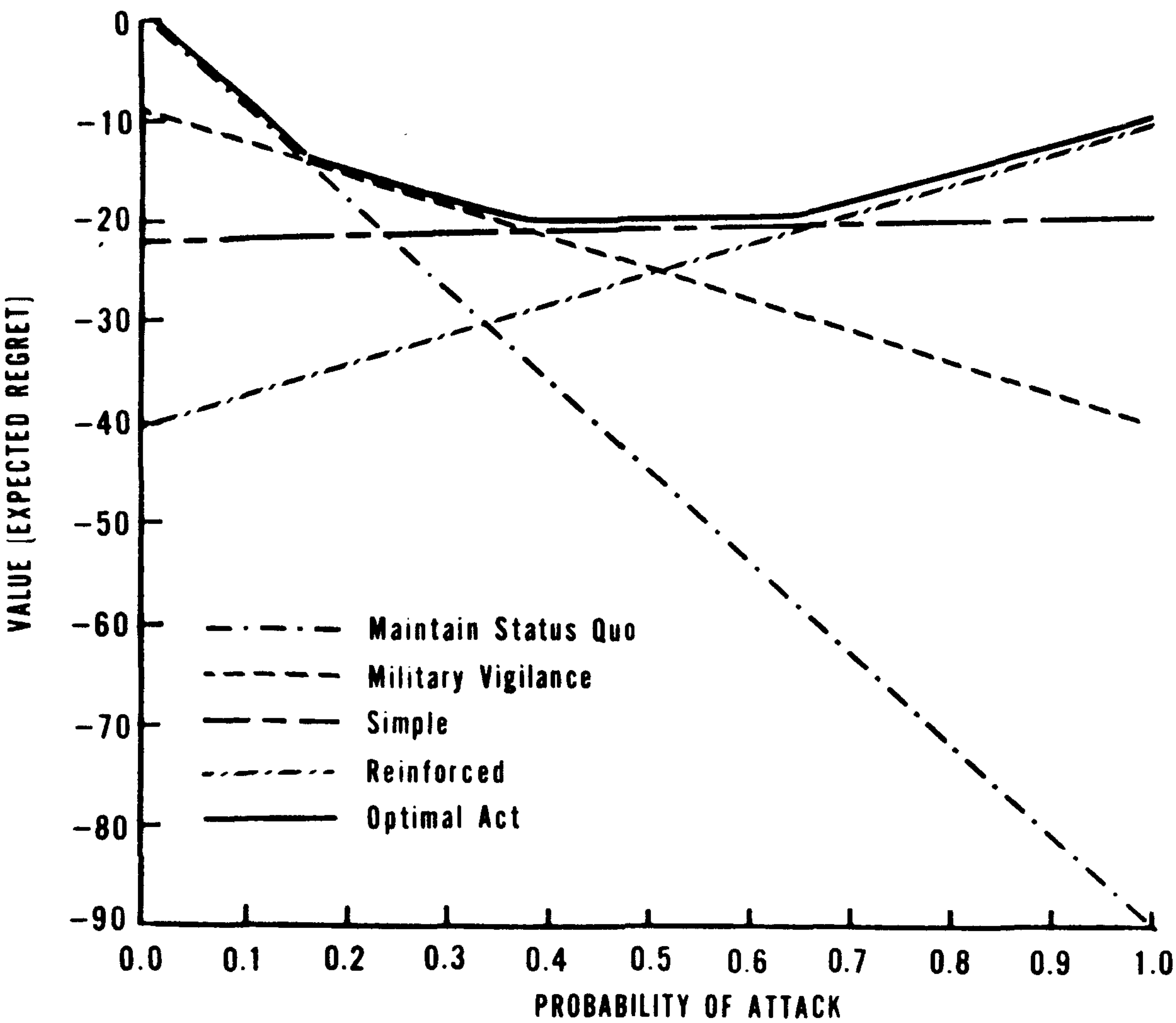


Figure A6-2. Changes in Expected Regret as a Function of Probability of Attack: Graphic Presentation

attack is simply one minus the probability of attack.) The expected regret scale runs from largest to smallest so that the smallest (most desirable) values will be at the top of the graph. An inspection of this graph enables the assessor to see at a glance the effect of changes in the probability assessment upon the choice of the optimal option. The option whose expected regret line is uppermost is the optimal act. In this example from the Warsaw Pact exercise, the status quo option is optimal until the probability of attack reaches about 0.17; at that point, military vigilance becomes optimal and remains so until the attack probability exceeds about 0.38, when simple alert becomes the option whose expected regret is smallest. The option of reinforced alert does not become optimal until the attack probability reaches 0.67. These points at which there is a shift in the optimal act are referred to as the "thresholds" of the probabilities.

(2) **Values and importance weights** - Holding the probability of attack constant (at the initial value of 0.10, for example), the user can also test the sensitivity of the output to changes in the importance weights, which will change the combined value matrix, and even to individual changes within any of the regret matrices for the different value dimensions. The first step would be to change the values of assigned weights, then recompute the combined value matrix as described in Table A6-8. Finally, revised expected regrets for each option are computed as shown in Table A6-9.

4. EVALUATION OF THE RAPID SCREENING METHOD. The usefulness of this method for the rapid screening of decision options can be judged by considering its various strengths and weaknesses. The main strengths of the procedure stem from the virtures generally claimed by decision analysis—quantification, normativeness, and communicability—all incorporated in a procedure that, compared to a full decision analysis, is relatively simple and rapid. Some of the weaknesses of the method, however, can also be attributed to this simplification which, at worst, may make the problem solved by the analysis so different from the actual decision problem faced that the solution is of no practical value.

a. **Strengths.** Like the standard decision-analytic procedure, this method for the rapid screening of decision options requires that the decision maker systematically list all decision options and event outcomes and express quantitatively the probability of occurrence for each outcome and the value of the different outcomes on several dimensions. This information is then processed mathematically to determine the optimal decision option and, through a sensitivity analysis, to reveal the assumptions and assessments which are critical to the choice of the best decision option. Such a formal procedure for decision making under uncertainty is generally considered to be superior to more intuitive methods where some factors may be overlooked or incorrectly weighted when their importance to the final decision is considered. Besides promising on the average and in the long run to give better decision making, the rapid screening procedure also promotes understanding of the problem both for an individual decision maker and within a group of decision makers. This increase in understanding occurs because the factors or events having an effect upon the probabilities must be enumerated during the probability assessment and because the dimensions and importance of the outcome values must be made explicit. Communication is improved among the people who are party to the decision; people with differing opinions can test the effect of their ideas on the final output; and, consequently, everyone's confidence in the correctness of (or at least the justification for) the selected decision option should be high.

The points cited above show how decision making for a particular problem may be improved by using the rapid screening method. In addition, the method has some usefulness as an introduction to the concepts of decision analysis and as a training device in the application of these concepts. A user's experience with one problem may in this way help to make the solution of future problems better and easier.

b. **Weaknesses**

(1) **Simplifications** - The simplified format of the rapid screening method differs from the standard decision analysis format in that (1) only one decision is allowed, followed by only one event node and (2) the probabilities of the different event outcomes are always independent of the action taken. If these implications are too restrictive, then the solution to the simplified problem (its best decision option) may not be a good approximation to the solution to the real (unsimplified) problem. For example, in the Warsaw Pact attack case, the assumption is made that the probabilities of attack versus no attack (initially assessed as 0.10 and 0.90) are independent of the decision option taken by NATO. In other words, whether NATO maintains the status quo or goes so far as to order a reinforced alert will have no effect upon the Pact's decision to attack or not. (Our interpretation of the 0.10 and 0.90 probabilities is then as follows: the Pact has already decided either to attack or not. NATO actions (within the range of options considered) will not change its decision. NATO does not know what the Pact's decision is but believes that there is a 0.10 probability that the intention is to attack and a 0.90 probability that it is not to attack.) If this assumption of independence of probabilities is incorrect, then the expected values of the regret under different decision options should be obtained by multiplying the values in the combined valuation regret matrix by different probabilities (newly assessed to account for the dependency) depending on which row (option) is being considered. This change in the calculation would generally result in different expected values and, consequently, might cause the optimal option (defined as the option having the smallest expected regret) to change.

The requirement of the simplified format that the problem has only one immediate decision node and one uncertain event node has the effect of eliminating the ability of the analysis to represent accurately a problem where there may be a sequence of decisions to be made. In the Warsaw Pact attack example, for instance, the probability graph (Figure A6-2) shows which decision is optimal for all possible attack probabilities from 0.0 to 1.0. If the probability is 0.50, for example, then the optimal decision is simple alert. It does not necessarily follow, however, that, if NATO actually went first to military vigilance on the basis of some data leading the probability of attack to be assessed at 0.30 (say) and then subsequently received information leading the probability to be revised to 0.50, simple alert would still be the optimal decision. This is because the regret matrices showing the values on different dimensions of the various options were assessed with the implicit assumption that the current status was no extraordinary alert position. If the status quo were simple alert, these values or their weights might be different.

(2) **Value assessments: payoff versus regret** - Another possible weakness of the method is that people will have a great deal of difficulty in assessing the outcome values on the artificial scales used here. If these assessments are not done coherently, then the output of the entire analysis is called into question.

In the examples shown here, the value within each dimension of a particular combination of decision option and event outcome was assessed, not by comparing all combinations to each other (payoff measures) on some absolute scale, but by taking each event outcome separately and, within that outcome column, assessing the regret resulting from making a non-optimal decision compared with the best possible decision under the outcome. One of the reasons for this approach is that, when the values of the options under one outcome are clustered at one end of the payoff scale and those under another outcome at the opposite end, the assessor may have difficulty in discriminating among the points in each cluster. For example, with two outcomes and four options, the payoff values assessed might be 0, 1, 2, 3 under one outcome and 100, 99, 98, 97 under the other. This lack of discrimination within a column is overcome by the technique of regret assessment which emphasizes comparisons within a column. Another reason for assessing regrets is that some assessors find it quite easy to answer questions phrased in regret terms (e.g., "Is it a bigger mistake in political cost terms to have failed to go to reinforced alert if there is an attack or to have mistakenly gone on reinforced alert when no attack occurs?"

However, despite these advantages of the regret assessment, this method may have some drawbacks. It may be that assessors have difficulty in keeping in mind what is meant by the regret measurement (namely, the comparison of a considered option to the optimal option) when using values of one column as a basis for getting those of another column or, what may be even more difficult, when comparing a column in one dimension to a column in another dimension to determine importance weights. The difficulty anticipated here is that an assessor will not be able to keep in mind simultaneously the three or four necessary factors. For intra-matrix comparisons, these factors are the optimal option and the considered option under one outcome versus the optimal option and the considered option under another outcome. For inter-matrix comparisons, the factors which may differ are the optimal option in each matrix, the considered option in each matrix, the outcome considered in each matrix, and the dimension of value. An assessor who has difficulty dealing with this complexity may initially assess values in terms of regret but then treat these as if they were payoffs in later stages of the assessment. For example, after the expert has assessed the regret matrices for military risk and for political cost in the Warsaw Pact attack exercise, he is asked, "Which is a worse mistake (and how much worse), -100 under military risk or -100 under political cost?" *Rather than considering this question in regret* terms, where "mistake" means regret at having chosen the wrong option when you could have chosen the optimal one, the assessor may respond on the basis of payoff, as if the question were, "Which option-outcome combination is worse (and how much), the military risk of an attack when you are in status quo readiness or the political cost under the same circumstances?"
Three possible ways of testing for the existence of this problem and overcoming the confusion are:

- o Assessing the value matrices and their importance weights both in terms of payoffs and of regrets. These assessments would be made at separate times and the results compared by looking at the regrets assessed directly and those computed from the payoffs;
- o Presenting the questions used to elicit the regret *assessments* as *paired* comparisons, without displaying the whole matrix to the assessor; and
- o Asking the assessor to justify each regret assessment with a few sentences explaining why one mistake is comparable to, or a certain amount worse than, another. By listening to these explanations, the elicitor may be able to tell whether the assessor is correctly considering the regret value rather than payoffs.

For the regret assessments presented in the examples of this paper, the third of these approaches was tried.

c. **Conclusions.** The overall experience with this simplified approach to the rapid screening of decision options is quite positive. The solutions to the problems to which it has been applied are seen as plausible by the users of the method in light of their explicit probability *and value assessments.* Furthermore, the discussion of these probabilities and values has improved communication among different parties to the decision. The users are also enthusiastic about their ability to modify by themselves both the structure of the problem and its inputs.

PROCEDURES USED IN PREPARING TABLE 5-15

1. Preparation of Verification Matrices. Table A7-1 illustrates how individual matrix values are obtained for the first matrix (P≥ 2%, P<2%) in Table 5-15. Values labeled, "a," "c," "a+b," and "c+d" are extracted from Part B, Table 5-14 and entered in the respective matrix positions (Table A7-1 and 5-15). Values for "a+c" and "a+b+c+d" are also obtained

Table A7-1. Example Computations Used to Prepare Matrices in Table 5-15.

EVENT OCCURRED	FORECAST PROBABILITY		TOTAL
	P ≥ 2%	P < 2%	
YES	(a) 293	(c) 19	(a+c) 312
NO	(b) 1008	(d) 888	(b+d) 1896
TOTAL	(a+b) 1301	(c+d) 907	(a+b+c+d) 2208

from Table 5-14, since they are the total numbers of events occurrences and forecasts, respectively. The three remaining values needed ("b," "d," and "b+d") are determined by calculating the differences between those values previously obtained. Similar procedures are used in preparing the other matrices.

2. Calculation of Post Agreement and Prefigurance (Panofsky and Brier, 1965).

a. Post Agreement. This is a measure of the reliability of categorical forecasts which describes the extent to which subsequent observations confirm the prediction, when a certain event is forecast. It indicates how frequently an event occurs when it was forecast. Table A7-2 shows the computations for the first matrix (P≥2%, P<2%) in Table 5-15. Notations (a, b, c, and d) and matrix values are taken from Table A7-1 above. Similar procedures are used in computing post agreement for all other matrices.

Table A7-2. Example Computation of Post Agreement.

EVENT OCCURRED	FORECAST PROBABILITY	
	P ≥ 2%	P < 2%
YES	$\frac{a}{a+b} = \frac{293}{1301} = 22.5\%$	$\frac{c}{c+d} = \frac{19}{907} = 2.1\%$
NO	$\frac{b}{a+b} = \frac{1008}{1301} = 77.5\%$	$\frac{d}{c+d} = \frac{888}{907} = 97.9\%$
TOTAL	100%	100%

b. Prefigurance. This is a measure of categorical forecasting capability which describes the extent to which the forecasts give advance notice of the occurrence of a certain event. It indicates how often an event is forecast when it occurs. Table A7-3 shows the computations for the first matrix ($P \geq 2\%$, $P < 2\%$) in Table 5-15. Notations and matrix values are obtained as stated above. Similar procedures are used in computing prefigurance for all other matrices.

Table A7-3. Example Computation of Prefigurance.

EVENT OCCURRED	FORECAST PROBABILITY		TOTAL
	$P \geq 2\%$	$P < 2\%$	
YES	$\frac{a}{a+c} = \frac{293}{312} = 93.9\%$	$\frac{c}{a+c} = \frac{19}{312} = 6.1\%$	100%
NO	$\frac{b}{b+d} = \frac{1008}{1896} = 53.2\%$	$\frac{d}{b+d} = \frac{888}{1896} = 46.8\%$	100%

INTRODUCTORY TRAINING SCENARIO

1. Background.

a. The text describes all the tools needed by skilled forecasters for producing good probability forecasts. The key element lacking is experience with this new approach. The limited experience resulting from a unit training program does not make a substantial difference in sharpness, since this attribute is similar in categorical forecasting and is dependent upon forecasting skill. However, reliability is a new ability that can be gained in a relatively short training environment. Experience by NWS shows that forecasters learn to adjust reliability biases very quickly, given timely feedback (Hughes, 1976a). Further, the experience within AWS indicates that reasonable reliability can usually be attained after a forecaster has issued 50-100 forecasts in which the event occurs. This does not mean that operational forecasts cannot be issued, when a forecaster has less experience. It simply means that the forecasts may not be as reliable as they could be.

b. This attachment provides guidance for establishing local training programs in probability forecasting. Two types of programs are needed. The first involves training forecasters who have no previous experience with probabilities. It must cover all phases of the effort. Forecasters completing this training should be fully capable of issuing reliable forecasts for the weather event used in training. The second program must train all forecasters to issue reliable forecasts for each weather event used operationally. Its objective is to provide forecasters with sufficient experience to establish reliability for that specific event. If time allows, experience can be gained by preparing training forecasts on a real-time basis. This if often not practical, especially when the forecasts are made only once a day. Further, it may take months or years to obtain adequate experience, where infrequent or rare weather events are concerned. Therefore, canned data, as described in this attachment, can be used to reduce training time considerably.

2. Preparation of the Training Programs.

a. Define the Event. The first step is to precisely define the weather event to be forecast. It must specify the data base time to be used in preparing the forecast, what element will be forecast, and when the forecast will be valid.

(1) Introductory training: The event chosen should have a climatic frequency of near 30%; be limited to a two category forecast; and have a lead time of approximately six hours. Note that rare events usually do not include a sufficient number of occurrences to gain experience or to perform reliable evaluations.

(2) Operational training. Train with an event as close as possible to the one that will be used operationally. The actual choice will be limited by the data base available.

b. Collect Data Base.

(1) Introductory training.

(a) Charts. Collect two sets of charts for 31 consecutive days each. A complete data base is not needed, because the principles can be learned from a minimum of data. Even daily surface charts, such as the US Dept of Commerce Daily Weather Map will suffice. The forecasts need not be for the home station. If available, choose charts for two different years (e.g., May 1976 and May 1977). If only local charts are used, two consecutive months in the same season are adequate, provided the forecasting techniques used are similar.

(b) Observations. Observations corresponding to the map times are needed. Verifying observations are also required for evaluating and critiquing the forecasts.

(c) Climatology. Obtain the best source of long-term climatology valid for the verifying time and give it to the forecasters. Use CC as a starting point for each forecast, if available.

(2) Operational training. A larger data base is needed for operational training to make the situation as realistic as possible. The extent of the data base is dictated by manageability. The type of data provided is also governed by the nature of the event. For rare events, it may be necessary to acquire several years of data for selected seasons in order to obtain sufficient forecasting experience.

c. Design of Worksheet and Verification Procedures.

(1) Introductory training. Design a worksheet similar to Figure 4-3 for recording and evaluating the forecasts. Instead of using zeros and ones to indicate the verification, enter the valid dates of the forecasts in the blocks on the forecast distribution diagram, and verify (occurrences of the event) by indicating slashes through the appropriate dates. Probability intervals must not be less than 20%, unless a large number of forecasts are used. Otherwise, the number of cases in the probability intervals may be too small to obtain reliable evaluations.

(2) Operational training. The worksheet format varies with the number of categories in the forecast. For a two category forecast, use the format given in Figure 4-3. For a larger number of categories, a form similar to the one described in Figure A8-1 is probably more suitable for recording the forecasts. Evaluations then take the format of Table 4-5. Use probability intervals for forecasts and evaluations which correspond to those used operationally, if known.

3. Training Procedures.

a. Introductory Training.

(1) Start the training program with a seminar covering all the key elements of probability forecasting, evaluation techniques, and a few examples of how probabilities are applied operationally.

(2) Follow with a workshop amplifying the basic concepts. Discuss the sequence of events to follow. Provide climatological aids, initial observations, charts, worksheets, etc, and instructions for completing the various tasks. Begin practice by having the trainees prepare probability forecasts for the chosen event using one of the two sets of data. Normally, an average of one minute per forecast is sufficient time for this phase.

(3) Give the trainees the observations to verify the practice forecasts. Have them compute, for each probability interval, the number of forecasts, number of event occurrences, observed frequency, and bias.

Compute Brier Scores if they are used routinely.

(4) Critique each trainee's forecasts. Discuss their merits and deficiencies, and how to overcome the biases. Cover sharpness, reliability, and the biases of over-underforecasting and over-underconfidence. If Brier Score measures are used, compare the individual scores with climatological scores.

(5) Use the second set of data to prepare another set of practice forecasts. This time, have the trainees concentrate on correcting the biases they had in the first set.

(6) Evaluate and critique this second set of forecasts. Most trainees achieve substantial improvement on the second set, but some will overcompensate (i.e., go from overforecasting to underforecasting). This simply means that they know the right principle, but need to achieve the right balance.

(7) After the workshop, require trainees to read this pamphlet to help reinforce the principles taught during the exercise, and to learn the details that will be needed before they can become proficient.

(8) Additional experience can be obtained by having the trainee issue practice forecasts on a real-time basis, or by using canned data. This is a good time to change to an event for which operational forecasts will be issued. Continue practice forecasts until the trainee attains the required reliability and sharpness.

b. Operational Training. Once introductory training is completed, it need not be repeated, except when the principles are not understood. Administer training on a real-time basis or by using canned data. Ideally, at least part of the training must be with real-time data, to better assess expected performance.

Year	Month	Day	Hour	Station Call Letters	Current Cig CAT	Current Vis CAT	Forecast Projection	CC Prob of Cig<200 (CAT A)	CC Prob of 200≤Cig<1000 (CAT B)	CC Prob of 1000≤Cig<3000 (CAT C)	CC Prob of 3000≤Cig (CAT D)	CC Prob of Vis<½ (CAT A)	CC Prob of ½≤Vis<2 (CAT B)	CC Prob of 2≤Vis<3 (CAT C)	CC Prob of 3≤Vis (CAT D)	Fcst Prob of Cig<200 (CAT A)	Fcst Prob of 200≤Cig<1000 (CAT B)	Fcst Prob of 1000≤Cig<3000 (CAT C)	Fcst Prob of 3000≤Cig (CAT D)	Fcst Prob of Vis<½ (CAT A)	Fcst Prob of ½≤Vis<2 (CAT B)	Fcst Prob of 2≤Vis<3 (CAT C)	Fcst Prob of 3≤Vis (CAT D)	Cig CAT at Valid Time	Vis CAT at Valid Time
77	10	01	00	FFO	D	D	3	2	6	10	82	1	3	6	90	8	7	7	78	5	10	10	75	D	D
77	10	01	00	FFO	D	D	6	4	6	12	78	3	5	10	82	12	10	10	68	10	13	13	64	D	C
77	10	01	06	FFO	D	C	3	5	7	13	75	8	27	32	33	13	15	15	57	15	15	25	45	D	C
77	10	01	06	FFO	D	C	6	5	8	15	72	7	22	30	41	10	12	15	63	12	13	20	55	D	D

I N S T R U C T I O N S

EXAMPLE: Shown above are a 3- and a 6-hour forecast made at 00Z and a 3- and a 6-hour forecast made at 06Z.

TIME FCST MADE (Z)	-	The transmission time(Z) of the forecast.
CURRENT CIG CAT	-	The AWS Cig category (A, B, C, D) at transmission time.
CURRENT VIS CAT	-	The AWS Vis category (A, B, C, D) at transmission time.
FORECAST PROJECTION	-	This added to the transmission time gives the valid time of the forecast (3=3hr fcst and 6=6hr fcst).
CC PROB OF ( )	-	The conditional climatological probability (%) of a specific Cig or Vis category ( ). Sum of probabilities for CAT A, B, C, D must equal 100.
CIG CAT AT VALID TIME	-	The Cig category that occurred at the valid time of the forecast.
VIS CAT AT VALID TIME	-	The Vis category that occurred at the valid time of the forecast.

Figure A8-1. Example Format for Recording Forecasts with More than Two Categories.

INTRODUCTION TO DECISION TREES

Decision trees are frequently used instead of decision matrices to solve problems. Most of the information was extracted from Selvidge's Technical Report 76-12, Rapid Screening of Decision Options (1976).

1. **General.**

a. "Decision Analysis" is the name given to a recently developed, formal procedure for resolving complex problems where the decision maker must choose from among a number of options, and where the best decision depends, in part, on some uncertain future events whose outcomes can only be guessed at when the decision is made.[1] The techniques of decision-analytic procedure help the decision maker to enumerate all the possible acts (called the decision options), and all the relevant uncertain events with their different possible outcomes. The procedure also requires the decision maker to express in numerical terms his feelings about the relative likelihood (called the probabilities) of different outcomes in conjunction with the different possible decision options. Once the decision problem has been described in this fashion, the decision-analytic procedure specifies the way in which this numerical information is aggregated into summary figures (one for each decision option). These are used as an indicator of the best decision option.

b. The description of a decision problem is generally presented in the form of a decision diagram, called a "decision tree," shown in Figure A9-1. In this format, decision points (called nodes) are represented by small squares, with the different possible options shown as lines or paths coming out of the square. Points or nodes where uncertain events occur are represented by small circles, with lines extending out to indicate the different possible outcomes of the event. One function of the decision tree is to illustrate how the decision problem unfolds over time. The decision and event nodes are arranged sequentially, in the order in which decisions must be made, and in which outcomes of the uncertain events are revealed to the decision maker.

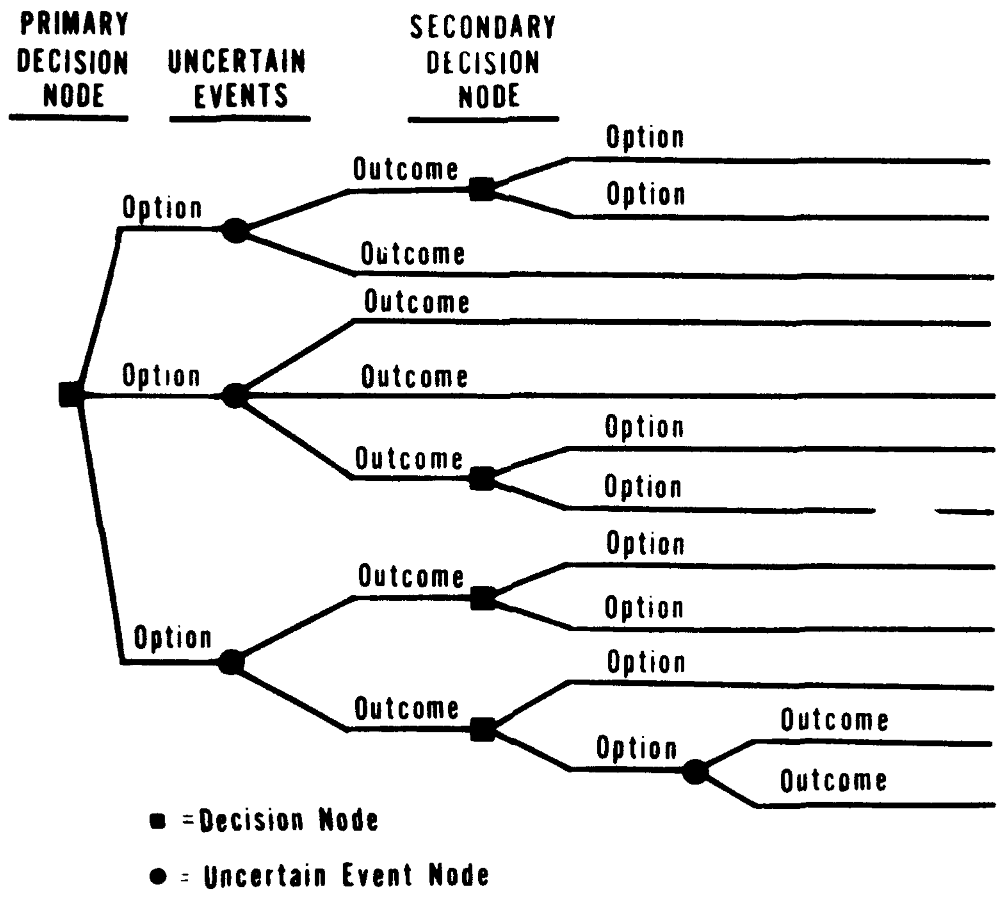


Figure A9-1. A Schematic of the Decision Tree Format.

[1]An excellent text on decision theory is Howard Raiffa's "Decision Analysis, Introductory Lectures on Choices under Uncertainty," Addison-Wesley Publishing Co, Reading, MA.

2. **Simplified Format Using a Decision Tree.** Figure A9-2 presents a simplified decision analysis format, showing a single decision node followed by a single uncertain event node. The probabilities of the different outcomes of the uncertian event (in this case, three outcomes are shown) are the same, regardless of which decision option is taken. Any end-point position of this simplified tree may be valued on many dimensions, and then summarized into a single utility figure.

3. ***Basic Matrix Format.***

a. The simplified decision tree contains all the information needed to carry out an approximate analysis of the problem. Since there is only one decision node and one uncertain event, an alternative way of displaying this information is in the form of a table or matrix. The rows represent the alternative decision options, and the columns of the matrix represent the different possible outcomes of the uncertain event. Each cell represents an option-outcome combination (and corresponds to an end-point in the decision tree). The cells contain the value of the particular option-outcome combination. There is a separate matrix for each value dimension.

b. Figure A9-3 shows how the decision sketched in Figure A9-2 appears in the basic matrix format.

c. The principal advantage of presenting decision problems in the basic matrix format is that people inexperienced in decision analysis seem to understand the matrix presentation more easily than the decision tree format. Additionally, the matrix provides a convenient way for recording the costs and benefits, when these need to be measured simultaneously in terms of a number of different factors (e.g., dollars, human lives, military advantage, political implications).

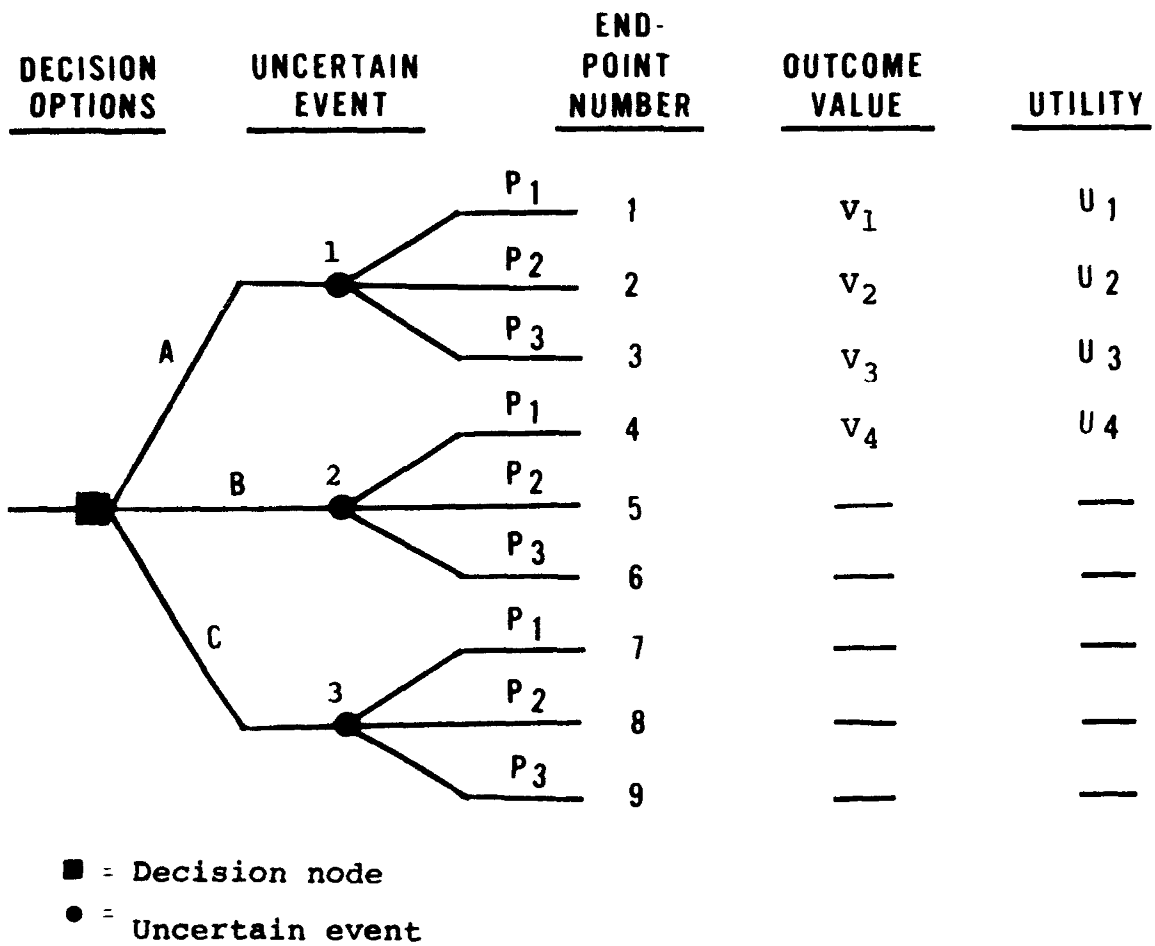


Figure A9-2. Simplified decision analysis format.

Uncertain Event

Decision Option

	1	2	3
A	V_1	V_2	V_3
B	V_4	V_5	V_6
C	V_7	V_8	V_9
	P_1	P_2	P_3

Event Probabilities

Figure A9-3. Basic Decision Matrix for the Decision Tree in Figure A9-2.

Instruction for Variable-Width Interval Forecasting of Maximum and Minimum Temperature

In forecasting the maximum (max) and minimum (min) temperature, you undoubtedly are somewhat uncertain about what the actual max and min will be. It is possible to give a point forecast (i.e., a single value) that represents your "best estimate" about the max or min, but the point forecast alone does not completely represent your uncertainty. A convenient way to convey this uncertainty is through the use of interval forecasts (i.e., intervals of values, as opposed to the single values used as point forecasts). Specifying an interval and the probability that the max (or min) temperature will be within the interval conveys a considerable amount of information about your uncertainty. On some days, you may feel that the odds are even that the max will be in a particular five degree interval; on other days, you may be much more uncertain, so you feel that the odds are even that the max will be in a particular **ten** degree interval. In this experiment you will be asked to determine an interval such that the probability is 50% that the max (or min) temperature will be in the interval, and you will be asked to determine an interval such that the probability is 75% that the max (or min) temperature will be in the interval. An interval is assumed to include its end points; for example, the interval 72-76°F is a five degree interval (it includes 72, 73, 74, 75, and 76). Note that in determining your interval forecasts, you will be working with intervals that are of **fixed probability** (50% and 75%), and you will have to determine the end points of the intervals; hence, the intervals are of **variable width** (the width depending on how uncertain you are on a given occasion).

The first step in determining the interval forecasts is to determine a median, which will be used as a mid-point for the variable width intervals. A median is a value that you feel is equally likely to be exceeded or not exceeded. For example, if you feel that it is equally likely that the max temperature tomorrow will be above 74 or below 74, then 74 is your median. The following dialogue should illustrate how you might arrive at a median.

Experimenter:	What is your best intuitive estimate of tomorrow's max temperature?
Forecaster:	About 90 degrees.
Experimenter:	My first step will be an attempt to sharpen up that initial estimate. If we were both to wager the same amount of money, would you rather bet that the max temperature will be above 90 degrees or below?
Forecaster:	Above 90 degrees.
Experimenter:	Would you rather bet that it will be above 94 degrees or below?
Forecaster:	Below.
Experimenter:	Above or below 91 degrees?
Forecaster:	Hmmm ... probably above.
Experimenter:	Above or below 92 degrees?
Forecaster:	It doesn't make much difference there.
Experimenter:	Above or below 93 degrees?
Forecaster:	Below.
Experimenter:	Fine. Then we will select 92 degrees as your indifference judgment. You think that is is just as likely that tomorrow's max temperature will be above 92 degrees as that it will be below 92 degrees. Is that right?
Forecaster:	That seems right.
Experimenter:	In a sense, 92 degrees, which is a median, is your best estimate of tomorrow's max temperature - it can be viewed as a point forecast.

The next step is to determine your 25th percentile (the median is sometimes called the 50th percentile). The 25th percentile is the value that divides the interval **below the median** into two equally likely subintervals. Note that the median divided the entire set of possible values into two equally likely intervals, so the procedure for determining the 25th percentile is very similar to the procedure for determining the median. For example, suppose that your median for the max temperature tomorrow is 74. Then if you feel that it is equally likely that the max temperature tomorrow will be below 71 or between 71 and 74, then 71 is your 25th percentile. The following continuation of the dialogue presented above illustrates the determination of a 25th percentile.

Experimenter:	In a sense, 92 degrees, which is a "median," is your best estimate of tomorrow's max temperature. The next series of questions that I'll ask is designed to explore just how certain you are that tomorrow's max temperature will be near 92 degrees. First, assume that all bets are off in case the max temperature is greater than 92 degrees. Do you think that it is more likely that tomorrow's max temperature will fall below 80 degrees or between 80 and 92 degrees? I am after two equally likely intervals below 92 degrees.
Forecaster:	It is more likely to be between 80 and 92 degrees.
Experimenter:	Below 85 degrees, or between 85 and 92 degrees?
Forecaster:	That's pretty difficult. Probably below 85 degrees.

Experimenter:	Below 84 degrees or between 84 and 92 degrees?
Forecaster:	That's about it. I can't choose between the two intervals.
Experimenter:	Fine - then we will accept 84 degrees as your 25th percentile.

Next, it is necessary to go through this type of procedure once more on the "low" side (the side below the median), in order to determine your 12½th percentile. As you can probably guess by now, the 12½th percentile divides the interval **below the 25th percentile** into two equally likely subintervals. The dialogue continues:

Experimenter:	Now that you've decided that 84 is your 25th percentile, let's assume that all bets are off if tomorrow's max temperature is above 84 degrees. Do you think that is is more likely that tomorrow's max temperature will fall below 70 degrees or between 70 and 84 degrees?
Forecaster:	Between 70 and 84 degrees.
Experimenter:	Below 75 degrees or between 75 and 84 degrees?
Forecaster:	Between 75 and 84 degrees.
Experimenter:	Below 80 degrees or between 80 and 84 degrees?
Forecaster:	That's pretty close, but I'd say below 80 degrees.
Experimenter:	Below 78 degrees or between 78 and 84 degrees?
Forecaster:	Between 78 and 84 degrees, but it's pretty close again.
Experimenter:	Below 79 degrees or between 79 and 84 degrees?
Forecaster:	I guess those intervals are about equally likely.
Experimenter:	Then we will select 79 degrees as your 12½th percentile.

The next step is to determine your 75th percentile, the value that divides the interval **above the median** into two equally likely subintervals. As you might suspect, the procedure for determining the 75th percentile is like the procedure for determining the 25th percentile. Let's go back to the dialogue.

Experimenter:	Now let's move on to the upper range, the range above the median. Assuming that all bets are off if tomorrow's max temperature is below 92 degrees, do you think that it is more likely to be between 92 and 100 or above 100?
Forecaster:	Definitely between 92 and 100.
Experimenter:	Between 92 and 95 or above 95?
Forecaster:	Still between 92 and 95.
Experimenter:	Between 92 and 94 or above 94?
Forecaster:	Now I am indifferent.
Experimenter:	In that case we will take 94 as your 75th percentile.

Finally, it is necessary to determine your 87½th percentile, the value that divides the interval **above the 75th percentile** into two equally likely subintervals. The procedure is similar to that for determining the 12½th percentile, so the dialogue might be as follows:

Experimenter:	If I can "push" you to determine one more indifference point, let's assume that all bets are off if the max temperature tomorrow is less than 94, which we just determined to be your 75th percentile. Do you think that the max temperature is more likely to be between 94 and 96 or above 96?
Forecaster:	Between 94 and 96.
Experimenter:	Between 94 and 95 or above 95?
Forecaster:	That's pretty difficult, but I guess I'm about indifferent.
Experimenter:	These are difficult judgments to make. Since you're about indifferent, we'll take 95 as you 87½th percentile.

The median, the 25th percentile, the 12½th percentile, the 75th percentile, and the 87½th percentile have been determined, in that order. These values can be used to determine interval forecasts. The probability is 50% that the max temperature will be between the 25th percentile and the 75th percentile, and the probability is 75% that the max temperature will be between the 12½th percentile and the 87½th percentile. Thus, we have one interval forecast with probability 50% and one with probability 75%. It is useful to reconsider the values that have been determined to make sure that they coincide with your best judgments. To illustrate this, we return to the dialogue one more time.

Experimenter:	Now let's carefully consider the values that you have estimated. First, consider the intervals A, B, C, and D, where A is below 84 degrees, B is between 84 and 92, C is between 92 and 94, and D is above 94. Assume that there is a four-way bet this time and you can pick only one of the intervals. Which one would you prefer?
Forecaster:	Hmmm ... Clearly not B or C. I guess I like A the best, but D looks pretty good, too.
Experimenter:	People occasionally squeeze the outside boundaries in too closely when making judgments like this for the first time.
Forecaster:	I must have done that because now I clearly like the outside two intervals better than the middle ones.
Experimenter:	Then move the outer boundaries out one degree each so that the boundaries are at 83 degrees, 92 degrees, and 95 degrees. Now which interval would you prefer to bet on?
Forecaster:	These estimates are better now. Any one of the intervals looks just as good as any other one to me. Also, I think that the max temperature is just as likely to fall inside the interval between 83 and 95 degrees as it is to fall outside that interval.
Experimenter:	Good. Now let's consider the interval's P, Q, R, and S, where P is below 79 degrees, Q is between 79 and 83, R is between 95 and 96, and S is above 96. I have taken the liberty of shifting your 87½th percentile up to 96, since the 75th percentile is now 95. In a four-way bet among these four intervals, which one would you prefer?
Forecaster:	The outside intervals look better again, so perhaps I need to move the 12½th and 87½th percentiles. Let's see - suppose they were 78 and 97. The 97 seems okay, but the 78 might still be a little high. I guess 77 and 97 would make me indifferent.
Experimenter:	Fine. Then your interval estimate with probability 50% is from 83 to 95, and your interval estimate with probability 75% is from 77 to 97. It is interesting that the boundaries are spread out asymmetrically around 92 degrees. The lower bound of 83 degrees has been pushed much farther away than the upper boundary of 95 degrees.
Forecaster:	I was thinking about that when making my estimates. A weak cold front is moving in from the northwest. It may reach here early tomorrow morning, but it may take until tomorrow night. If it gets here before morning, then it won't get very warm tomorrow. But, if the front is delayed, then the max temperature should be around 92 degrees.
Experimenter:	Then that explains why the upper boundary is so much closer to 92 degrees. There is little chance for any change in conditions to produce much of an increase above your median of 92.
Forecaster:	That's right. Looked at that way, these intervals display a lot of what I know about tomorrow's max temperature. They don't indicate why the max temperature could drop but they certainly show that it can. I wouldn't expect to always have such asymmetric intervals when compared with the median, but it sure seems reasonable in this particular situation.

For convenience, here is a summary of the procedure. First, consider the maximum temperature in degrees Fahrenheit (on the day shift, this refers to tomorrow's maximum; on the midnight shift, this refers to today's maximum) and complete the following steps:

1. Determine your median.
2. Determine your 25th percentile.
3. Determine your 12½th percentile.
4. Determine your 75th percentile.
5. Determine your 87½th percentile.
6. Look at the resulting intervals to make sure that they agree with your judgments, making any changes you deem necessary.

Next, consider the minimum temperature in degrees Fahrenheit (on both the day and midnight shifts, this refers to tonight's minimum), and repeat the six steps listed above.

DEPARTMENT OF THE AIR FORCE
Headquarters Air Weather Service (MAC)
Scott Air Force Base, Illinois 62225-5008

CHANGE 1
AWSP 105-51
16 October 1989

Weather

PROBABILITY FORECASTING: A GUIDE FOR FORECASTERS AND STAFF WEATHER OFFICERS

AWSP 105-51, 31 October 1978, is changed as follows:

Write-In Changes:

Page	Paragraph/Section	Line	Action
1-1	1-1	-	Delete first sentence.
1-1	1-1	6	Change "... to implement that policy." to "on probability forecasting."
1-1	1-4b, Example 1	-	Delete last sentence, "AWS is developing..."
2-3	2-4, Example 2	-	Delete last sentence.
4-1	4-1	-	Delete last sentence.
4-5	4-3	14	Change "indicater" to "indicate."
6-19	6-5b	5	Change "RUSSWOs" to "Surface Observation Climatic Summary (SOCSs)."
6-21	6-5b	-	Delete from "AWS is developing..." to end of paragraph.
6-23	6-5c(6)	5	Change "RUSSWOs" to "SOCSs."

OFFICIAL

JOHN J. KELLY, JR.
Brigadier General, USAF
Commander

JOHN C. EVANS, SR.
Information Management Officer

No. of Printed Pages: 1
Distribution: F;X