AN ABSTRACT OF THE THESIS OF

Wu-ron Hsu for the degree of Master of Science in Atmospheric Sciences presented on June 24, 1982.

Title: Evaluation of a Probabilistic Quantitative Precipitation Forecasting Experiment

Abstract approved: Redacted for Privacy

Forecasts of the likelihood of occurrence of various amounts of precipitation are very important, since excessive precipitation amounts over relatively short time periods can have adverse effects on public safety and economic efficiency. As a result, forecasters at the National Weather Service Forecast Office in San Antonio, Texas were asked to formulate subjective probabilistic quantitative precipitation forecasts on an experimental basis beginning in February 1981. This study describes methods of evaluating probability forecasts of this ordinal variable and presents some results of the first year of the experiment.
Scalar and vector evaluation procedures are described. In the case of scalar evaluation, the inclusion of a no-skill line and a no-correlation line on reliability diagrams is helpful in representing the skill, reliability, and resolution geometrically in two-state situations. Geometrical interpretations of attributes of forecasts can also be accomplished in three-state situations based on vector evaluation procedures. A skill score for subsample forecasts is shown to be useful in identifying systematic errors made by forecasters or forecast systems. A beta model is developed to obtain a forecaster's predictive distribution (i.e., the distribution of use of probability values).

The experimental results show that the skill of the subjective forecasts is generally higher than the skill of objective guidance forecasts for measurable precipitation (i.e., precipitation amounts exceeding a threshold of 0.01 inches), but that the opposite is true for thresholds associated with larger precipitation amounts. This result is due primarily to the forecaster's tendency to overforecast for the events associated with higher precipitation thresholds. The tendency to overforecast is most pronounced in the nighttime forecasts and in the forecasts for drier stations. The MCS objective guidance forecasts, on the other hand, are quite reliable for both
periods and all stations. The vector evaluation approach indicates that the degree of overforecasting is quite high for bimodal forecasts and that the skill contribution from bimodal forecasts is negative in many cases.
Evaluation of a Probabilistic Quantitative Precipitation Forecasting Experiment

by

Wu-ron Hsu

A Thesis
submitted to

Oregon State University

in partial fulfillment of the requirements for the degree of

Master of Science

Completed June 24, 1982
Commencement June 1983
I would like to express my deepest gratitude to Dr. Allan H. Murphy for his invaluable advice and guidance. His patience in correcting the English structure of my thesis made possible the presentation here. Sincere thanks goes to Dr. Richard W. Katz for his valuable comments on many theoretical parts of this study.

I would also like to express my appreciation to the forecasters at the WSFO in San Antonio, Texas for their cooperation and participation in the quantitative precipitation forecasting experiment. The assistance of Messrs. Gifford F. Ely, Jr. (Deputy MIC, San Antonio WSFO) and Daniel L. Smith (Chief, Scientific Services Division, NWS Southern Region Headquarters, Fort Worth, Texas) in the design and conduct of the experiment is gratefully acknowledged.

This work was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under grant ATM88-04689.
# TABLE OF CONTENTS

1  INTRODUCTION  

2  QUANTITATIVE PRECIPITATION FORECASTS  
2.1 Current Procedures and Practices  
2.2 The Probabilistic QPF Experiment  
2.3 Some Notation and Definitions  
2.3.1 Forecasts/Observations and Cumulative Forecasts/Observations  
2.3.2 Subsamples (Subsets) of Cumulative Forecasts/Observations  

3  THE RPS AND ATTRIBUTES OF PROBABILITY FORECASTS  
3.1 The Definition of the RPS  
3.2 Attributes and Measures of Attributes  
3.2.1 Definitions of Attributes  
3.2.2 Partitions of the RPS: Measures of Reliability and Resolution  
3.2.2.1 Two-Term Partition of the RPS  
3.2.2.2 Three-Term Partition of the RPS  
3.2.3 Measures of Skill  
3.2.3.1 Overall Skill  
3.2.3.2 Skill for Specific Subsample Forecasts  

Page numbers: 1, 4, 6, 8, 9, 11, 13, 14, 15, 17, 18
4 SOME THEORETICAL PROPERTIES OF THE RPS IN THE
TWO-STATE SITUATION
4.1 The RPS as the Variance of Accuracy
4.2 \text{Var}(R-D) for Several Special Kinds of
Forecasts
4.2.1 Forecasts and Observations Uncorrelated
4.2.2 Sample Climatological Forecasts and the
Definitin of \( s^* \)
4.2.3 Perfectly Reliable Forecasts
4.2.4 Perfect Forecasts
4.3 The Probability Distribution of R-D
4.3.1 The Means of R|D=0 and R|1|D=1
4.3.2 The Variances of R|D=0 and R|1|D=1
4.4 A Beta Model for R-D
4.5 An Example

5 METHODS OF EVALUATION
5.1 Scalar Methods
5.1.1 Geometrical Interpretation of the
Attributes of Probability Forecasts on a
Reliability Diagram
5.1.2 Test of Reliability
5.2 Vector Methods
5.2.1 A Geometrical Framework for the
Three-State RPS
5.2.2 Geometrical Interpretation of Attributes
of Probability Forecasts
LIST OF FIGURES

4.1. The reliability diagram. 26
4.2. The domain of R-O. 31
4.3. The distributions of R-O for $s_0=0$ and $s_1=1$. 33
4.4. The distributions of R of the MOS forecasts and the beta model. 37
4.5. The distributions of R of the SUB forecasts and the beta model. 38
4.6. The distributions of R-O of the MOS forecasts and the beta model. 39
4.7. The distributions of R-D of the SUB forecasts and the beta model. 40
5.1. Reliability diagram and the geometrical representation of resolution, reliability, and skill in the two-state situation. 44
5.2. Geometrical framework for the RPS in the three-state situation. 50
5.3. "Uncertainty" as a function of $\bar{Q}$. 54
5.4. Geometrical representation of resolution, reliability, and skill in the three-state situation. 56
6.1. Reliability diagram for forecasts of precipitation $\geq 0.01$ inches. 60
6.2. Reliability diagram for forecasts of precipitation $\geq 0.25$ inches. 61
6.3. Reliability diagram for forecasts of precipitation $\geq 0.50$ inches. 62
6.4. Reliability diagram for forecasts of precipitation $\geq 1.00$ inches. 63
6.5. Reliability diagram for forecasts of precipitation $\geq 2.00$ inches. 64
6.6. Distribution of $K_5$, $s_4$, and $sp_5$ for the MOS forecasts for the forecasts with a threshold of 0.01 inches. 69
6.7. Reliability diagram for the daytime forecasts of precipitation ≥ 0.25 inches.

6.8. Reliability diagram for the nighttime forecasts of precipitation ≥ 0.25 inches.

6.9. The distribution of $K_t$ and $s_{0_t}$ for the MOS forecasts in the 4-5-6 case.

6.10. The distribution of $K_t$ and $s_{0_t}$ for the MOS forecasts in the 2-5-6 case.

6.11. The $s_{0_t}$ for the nighttime SUE forecasts in the 4-5-6 case.

6.12. The density function of a bimodal forecast $R=(0.3, 0.3)$

6.13. $R_{1t} - \bar{D}_{1t}$ for the SUB forecasts in the 4-5-6 case.

6.14. $R_{2t} - \bar{D}_{2t}$ for the SUB forecasts in the 4-5-6 case.
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1.</td>
<td>Climatological probabilities and average forecast probabilities for MOS and SUB.</td>
<td>59</td>
</tr>
<tr>
<td>6.2.</td>
<td>The overall skill for the MOS and SUB forecasts for different time periods.</td>
<td>70</td>
</tr>
<tr>
<td>6.3.</td>
<td>The RPS, the terms in the partition of the RPS, and the skill scores for MOS, SUB, and CLI in the two-state situation.</td>
<td>74</td>
</tr>
<tr>
<td>6.4.</td>
<td>Overall skill for four stations.</td>
<td>77</td>
</tr>
<tr>
<td>6.5.</td>
<td>The RPS, the terms in the partition of the RPS, and the skill scores for the MOS, SUB, and CLI forecasts in the simplified three-state situation.</td>
<td>78</td>
</tr>
</tbody>
</table>
Evaluation of a Probabilistic Quantitative Precipitation Forecasting Experiment

CHAPTER 1

Introduction

Forecasts of the occurrence of measurable precipitation (≥0.01 inches) are made on an operational basis by National Weather Service (NWS) forecasters, and they are routinely disseminated to the general public and specific users (e.g., see Hughes, 1980). As an aid in preparing these subjective forecasts, the forecasters receive objective guidance forecasts of precipitation occurrence based on the model output statistics (MOS) system (Glahn and Lowry, 1972; Lowry and Glahn, 1976). Both the objective MOS forecasts and the subjective forecasts formulated by NWS forecasters are expressed in probabilistic terms. It has been demonstrated repeatedly that a probabilistic mode of expression significantly enhances the value and usefulness of weather forecasts.
Knowledge of the likelihood of occurrence of various amounts of precipitation is also very important, since excessive precipitation amounts over relatively short time periods can have adverse effects on public safety and economic efficiency. Forecasts of precipitation amounts are generally referred to as quantitative precipitation forecasts (QPFs; for convenience, we will also use the acronym QPF to represent "quantitative precipitation forecasting"). Unfortunately, accurate prediction of precipitation amounts is difficult and remains largely an unsolved meteorological problem (Committee on Atmospheric Sciences, 1983). For this and other reasons, weather forecasts currently formulated by NWS forecasters and disseminated to the public generally do not contain quantitative information concerning precipitation amounts. Moreover, when information related to precipitation amounts is included in such forecasts, it is usually expressed qualitatively in a relatively imprecise manner.

In view of the need for quantitative information concerning precipitation amounts, an experiment was undertaken to investigate the ability of NWS forecasters to formulate probabilistic QPFs. The details of this experiment and the current procedures concerning QPFs are
presented in Chapter 2. Chapter 3 introduces the ranked probability score (RPS), the scoring rule used to evaluate the QPFs in this experiment, and describes some relevant attributes of probability forecasts and their measures. Chapter 4 discusses some theoretical properties of the RPS in the two-state situation. Chapter 5 describes the evaluation methods used in this study. The experimental results are presented in Chapter 6, and a brief conclusion is contained in Chapter 7.
In this chapter, Section 2.1 briefly describes current NWS procedures and practices with regard to quantitative precipitation forecasting. The probabilistic quantitative precipitation forecasting experiment is discussed in Section 2.2. Section 2.3 introduces some basic notation and definitions that will be used extensively in later chapters.

2.1 Current Procedures and Practices

Routine subjective forecasting of precipitation amounts in the NWS began in 1961 with the establishment of a specialized group of meteorologists at the National
Meteorological Center (NMC). This group (the so-called QPF Branch of NMC) prepares QPFs for the United States up to two days in advance. The QPFs delineate areas where precipitation amounts are expected to equal or exceed certain threshold values (namely, 0.25, 0.50, 1.00, and 2.00 inches). These QPFs are categorical forecasts (i.e., the uncertainty inherent in the forecasts is not specified). To the author's knowledge, subjective probabilistic QPFs have not been prepared on a regular basis anywhere in the U.S.

The MOS QPFs became available in 1975. They are prepared primarily as guidance to the QPF Branch of NMC. There are currently two MOS probability of precipitation amount (POPA) systems in operation. The first uses predictors from the NMC's limited area fine mesh (LFM) model, whereas the second uses predictors from NMC's primitive equation (PE) model and the NWS Techniques Development Laboratory's trajectory model. Forecasts for four thresholds (≥0.25, ≥0.50, ≥1.00, ≥2.00 inches) are made for periods up to 60 hours in advance. A method of transforming probability forecasts into categorical forecasts has been used to make comparisons between the subjective categorical QPFs and the MOS QPFs (Bermowitz and Zurndorfer, 1979).
2.2 The Probabilistic QPF Experiment

Since only categorical subjective QPFs are available and a need exists for quantitative information concerning the uncertainty inherent in subjective forecasts of precipitation amount, an experiment was undertaken to investigate the ability of NWS forecasters to formulate QPFs in terms of probabilities. The experiment involved the formulation of probabilistic QPFs by forecasters at the San Antonio Weather Service Forecast Office. Specifically, these forecasts expressed the likelihood that certain threshold amounts of precipitation would be equaled or exceeded in 12-hour periods at four locations in Texas; namely, Austin (AUS), Brownsville (BRO), Houston (IAH), and San Antonio (SAT). The threshold amount values used in this experiment were 0.25 inches, 0.50 inches, 1.00 inches, and 2.00 inches. The permissible probability values (in %) included 0, 1, 2, ..., 99, 100. The experimental forecasts were made each day at approximately 0300 LST and 1500 LST for 12-hour periods from 0600-1800 LST (1200-2400 GMT) and 1800-0600 LST (0000-1200 GMT), respectively. We shall refer to the forecasts for the 1200-2400 GMT period as the daytime forecasts and to the forecasts for the 0000-1200 GMT period as the nighttime forecasts.
The forecasters who prepared the QPFs also made PoP (probability of measurable precipitation) forecasts for the same 12-hour periods at each location. As indicated previously, the MOS system produces both PoP forecasts and probabilistic QPFs, and these forecasts are prepared for the same threshold values, time periods, and locations as the subjective forecasts. For convenience, we will sometimes refer to these objective and subjective QPFs as the MOS and SUB forecasts, respectively.

The experiment was initiated in February 1981, after a short period of familiarization and practice forecasting. In this study, we report results based on the first 12 months of the experimental program (February 1981 through January 1982). During this period a total of 622 subjective QPFs were made for each location, including 339 daytime forecasts and 283 nighttime forecasts. Observations of precipitation amounts in the corresponding 12-hour periods at these locations were used to verify the objective and subjective forecasts.

Long-term monthly climatological probabilities for the 20-year period from 1960-1979 have been calculated for all five threshold values of precipitation amount for day and night periods at each location. These probabilities are taken to represent climatological forecasts and are
used to provide a standard of reference in computing the skill of the MOS and SUB forecasts. For convenience, we will call these forecasts the CLI forecasts.

2.3 Some Notation and Definitions

2.3.1 Forecasts/Observations and Cumulative Forecasts/Observations

The range of precipitation amount has been divided into a set of 6 mutually exclusive and collectively exhaustive states or events \( \{e_1, e_2, \ldots, e_6\} \), where \( e_1 \) represents the state of precipitation amount greater than 2.00 inches in a 12-hour period, \( e_2 \) represents the state of precipitation amount between 1.00 and 2.00 inches in a 12-hour period, \ldots, and \( e_6 \) represents the state of no precipitation or trace in a 12-hour period. The set of probabilities assigned to the 6 states on each occasion constitutes one forecast, and we denote this forecast by a row vector \( \mathbf{r}=(r_1, r_2, \ldots, r_6) \) \( (r_n \geq 0, \Sigma r_n = 1; \ n=1, \ldots, 6) \), where \( r_n \) is the probability assigned to state \( e_n \) on the occasion of concern. Similarly, we denote the relevant observation by a row vector \( \mathbf{d}=(d_1, d_2, \ldots, d_6) \), where \( d_i=1 \) if state \( e_i \) occurs and \( d_n=0 \) for \( n \neq i \).
In the QPF experiment, however, the forecasts are made cumulatively. We use the capital letter $R$ to denote the row vector of cumulative probability forecasts, in which the $n$th component is $R_n = \sum_{i=1}^{n} r_i$ \((n=1,\ldots,6)\). Since there are 5 thresholds (6 states) involved, each cumulative forecast is in the form $R = (R_1, R_2, R_3, R_4, R_5, 1)$, where $R_n$ is the forecast probability of precipitation amounts greater than some threshold value $x_n$ inches in a 12-hour period (in which $x_1=2.00$, $x_2=1.00$, $x_3=0.50$, $x_4=0.25$, and $x_5=0.01$). The row vector $D$, whose component $D_n$ is defined as $\sum_{i=1}^{n} d_i$ \((n=1,\ldots,6)\), represents the cumulative observation corresponding to the cumulative forecast $R$. For example, the cumulative forecast $R = (10\%, 20\%, 30\%, 40\%, 50\%, 100\%)$ corresponds to the noncumulative forecast $r = (10\%, 10\%, 10\%, 10\%, 10\%, 50\%)$. An observation of precipitation amount of 0.40 inches in a 12-hour period is represented by $d=(0,0,0,1,1,1)$, or by $d=(0,0,0,1,0,0)$.

2.3.2 Subsamples (Subsets) of Cumulative Forecasts/Observations

In evaluating the results of the probabilistic QPF experiment, each component in the (cumulative) forecast $R$ has been rounded to the nearest ten percent. Thus, only 11 distinct values \((0, 10\%, \ldots, 100\%)\) have been used as
probability values for the components in the (cumulative) forecast R.

In later discussions, the phrase "scalar evaluation procedure" will mean that a particular component of the cumulative forecast R (e.g., $R_3$) has been treated as an individual forecast. We can then identify 11 distinct "subsamples" (or subsets) of the entire sample of forecasts, and all the forecasts in a particular subsample will possess the same probability value.

For vector evaluation procedures, two (or more) components of the cumulative forecast R (e.g., $R_4$ and $R_5$ in $R = (R_4, R_5, 1)$; see Section 6.2) together are treated as an individual forecast. In these cases, we can identify at most $W$ distinct forecasts (or subsamples), where

$$W = \sum_{i=1}^{m} \binom{m+i-1}{i-1}(12-i)$$

(Murphy 1972a), in which $\binom{x}{y} = x!/(y!(x-y)!)$ for $x \geq y \geq 0$, $\binom{x}{y} = 1$ for $x = 0$ and $y = 0$, $\binom{x}{y} = 0$ otherwise, and $m$ is the (total) number of components involved. When $m=3$, $W=66$. 
CHAPTER 3

The RPS and Some Attributes of Probability Forecasts

In this chapter, Section 3.1 contains a definition of the ranked probability score (RPS) (Epstein, 1969). Section 3.2 describes some attributes of probability forecasts and demonstrates that the RPS and its partition provide measures of all of these attributes.

3.1 The Definition of the RPS

The RPS for a collection of K vector forecasts $\mathbf{R}_k = (R_{1k}, R_{2k}, \ldots, 1)$ and the K relevant vector observations $\mathbf{D}_k = (D_{1k}, D_{2k}, \ldots, 1)$ $(k=1, \ldots, K)$ is defined as

$$ \text{RPS} = \frac{1}{K} \sum_{k=1}^{K} (R_k - D_k)'(R_k - D_k), \quad (3.1) $$

(in which a prime denotes a column vector) or
\[ K = \frac{1}{10} \sum_{h=1}^{K} (R_{n_h} - D_{n_h})^2. \] (3.2)

In the two-state situation, the RPS can be expressed simply as

\[ \text{RPS} = \frac{1}{K} \sum_{h=1}^{K} (R_{n_h} - D_{n_h})^2, \] (3.3)

since \( R_{2h} = D_{2h} = 1 \) for all \( h \). The magnitude of the RPS in such situations is just one half of the Brier score (Brier, 1950).

We shall use the RPS as the basic scoring rule throughout this study for the following reasons: (1) the RPS is particularly well suited for the evaluation of forecasts of ordinal variables (such as precipitation amount), (2) the RPS is a measure of accuracy for forecasts of such variables, (3) the RPS is strictly proper (Murphy, 1970), and (4) its partitions provide measures of several attributes of probability forecasts.
3.2 Attributes and Measures of Attributes

3.2.1 Definitions of Attributes

Inferential attributes of probabilistic forecasts are concerned with the correspondence between forecasts and observations, either individually or collectively (Murphy, 1977). We shall be mainly concerned with four such attributes in this study: (1) accuracy, (2) reliability, (3) resolution, and (4) skill.

Accuracy is generally defined as the average degree of correspondence between individual forecasts and the relevant observations over a set of forecasts. Reliability refers in general to the correspondence between average forecasts and average observations (or observed relative frequencies). Resolution is concerned with a forecaster's ability to divide the set of K forecasts into subsets (1) for which the average observations approach, as closely as possible, vectors which represent the occurrence of one of the events (in other words, vectors with components that have values of either zero or one) or (2) for which the average observations corresponding to the subsets differ, as much
as possible, from the average observation for the entire set of forecasts. In this study, we will use the second definition of this attribute. Skill refers to the average accuracy of the forecasts relative to the average accuracy of forecasts produced by some reference procedure such as climatology.

3.2.2 Partitions of the RPS: Measures of Reliability and Resolution

3.2.2.1 Two-Term Partition of the RPS

Murphy (1972b) has shown that the RPS can be partitioned as follows:

\[
\text{RPS}(\bar{R}, \bar{Q}) = \frac{1}{K} \sum_{t=1}^{T} K_t (R_t - \bar{Q}_t) (R_t - \bar{Q}_t)^* + \frac{1}{K} \sum_{t=1}^{T} K_t \bar{Q}_t (U - \bar{Q}_t)^*
\]

(3.4)

where \( U \) is a row vector whose \( N \) components are all equal to one (i.e., \( U=(1, \ldots, 1) \)) and \( T \) is the number of distinct vector forecasts. Also, we define

\[
\bar{Q}_t = \frac{1}{K_t} \sum_{k=1}^{K_t} Q_{kt},
\]

where \( K_t \) is the sample size for the \( t \)-th subsample and \( Q_{kt} \) is the \( k \)-th of the \( K_t \) observation vectors in this subsample.
Note that both terms on the right-hand side of Eq. 3.4 are nonnegative. The first term is the sum of the squares of the differences between the average forecast \((\bar{R}_t)\) and the average observation \((\bar{O}_t)\) for all \(I\) subsamples. This quantity is clearly a reasonable measure of the attribute reliability defined in Section 3.2.1. We shall call this term the reliability term. The reliability of the \(t\)-th subsample refers to the magnitude of contribution of any one forecast in that subsample to the overall reliability term. Thus, \((R_t-\bar{O}_t)(R_t-\bar{O}_t)^*\) is the reliability of the \(t\)-th subsample.

The second term on the right-hand side of Eq. 3.4 is a measure of the attribute resolution according to its first definition in Section 3.2.1. However, the second definition of resolution is more useful than this first definition, and the former relates to the resolution term in a three-term partition of the RPS.

3.2.2.2 Three-Term Partition of the RPS

The three-term partition of the RPS can be expressed as follows:

\[
RPS(R, O) = \bar{O}(\bar{U} - \bar{O})^* + \frac{1}{K} \sum_{t=1}^{T} K_t (R_t - \bar{O}_t)(R_t - \bar{O}_t)^*
\]
\[ - \frac{1}{K} \sum_{t=1}^{T} \frac{1}{K} (\tilde{Q}_t - \bar{Q})(\tilde{Q}_t - \bar{Q}) \]  
(Murphy, 1973). The first term on the right-hand side is a measure of the uncertainty inherent in the events on the K occasions of concern. The larger this term, the greater is the RPS. But larger values of this term do not necessarily mean that the skill of the forecaster is worse. The uncertainty term depends neither on \( \bar{R}_t \) (forecast probability) nor on \( \bar{Q}_t \), but only on the overall relative frequencies of the events \( \bar{Q} \). This term is the variance of the observation vectors \( Q_k \) in the sample.

The second term on the right-hand side of Eq. 3.5 is the reliability term as in the two-term partition. The third term is a measure of resolution according to the second definition in Section 3.2.1. Since this term is always negative, we shall use the absolute value of this term to denote resolution. The resolution of the t-th subsample is defined as \( (\tilde{Q}_t - \bar{Q})(\tilde{Q}_t - \bar{Q})^* \). For convenience, we will denote the overall reliability as \( RE \) and the overall resolution as \( RS \).
3.2.3 Measures of Skill

3.2.3.1 Overall Skill

Measures of skill involve a comparison between the forecasts and some reference procedure (see Section 3.2.1). Here, besides using long-term climatology as the reference procedure, we will also use the so-called sample climatology as described in the previous chapter. Let \( RPS^{CL} \) denote the RPS for forecasts based on long-term climatological probabilities. The skill with respect to long-term climatology can then be defined as

\[
S^{CL} = \frac{RPS^{CL} - RPS}{RPS^{CL}}. \tag{3.6}
\]

Let \( RPS^C \) denote the RPS for forecasts based on sample climatological probabilities. Then, substituting \( R = \bar{d} \) into Eq. 3.5, we have

\[
RPS^C = \bar{d} (\bar{U} - \bar{d})^*, \tag{3.7}
\]

which is the uncertainty term in the three-term partition of the RPS. Thus, the skill with respect to sample climatology is defined as

\[
S = \frac{(RPS^C - RPS)}{RPS^C} = \frac{(RS - RE)/(\bar{d} (\bar{U} - \bar{d})^*)}{(RS - RE)/(\bar{d} (\bar{U} - \bar{d})^*)}, \tag{3.8}
\]

which is the difference between the overall reliability
and resolution divided by the uncertainty (or variance of the observations).

### 3.2.3.2 Skill for Specific Subsample Forecasts

For the constant forecast associated with a specific subsample, only the skill with respect to sample climatology is calculated in this study. Let \( RPS_t \) denote the RPS for the \( t \)-th subsample and let \( RPS^c_t \) denote the RPS for the sample climatological forecast for that subset. Using Eq. 3.4, we have

\[
RPS_t = RPS(B_t, \bar{Q}_t) = \frac{R_t - \bar{Q}_t}{\bar{Q}_t} \frac{(B_t - \bar{Q}_t)^* + \bar{Q}_t (\bar{Q}_t - \bar{Q}_t)^*}{\bar{Q}_t (\bar{Q}_t - \bar{Q}_t)^* + \bar{Q}_t (\bar{Q}_t - \bar{Q}_t)^*} \tag{3.9}
\]

and

\[
RPS^c_t = RPS(\bar{Q}_t, \bar{Q}_t) = \frac{(\bar{Q}_t - \bar{Q}_t)^* + \bar{Q}_t (\bar{Q}_t - \bar{Q}_t)^*}{\bar{Q}_t (\bar{Q}_t - \bar{Q}_t)^* + \bar{Q}_t (\bar{Q}_t - \bar{Q}_t)^*} \tag{3.10}
\]

The skill \( ss_t \) for the \( t \)-th subsample is then

\[
ss_t = \frac{(RPS^c_t - RPS_t)}{RPS^c_t} = \frac{(\bar{Q}_t - \bar{Q}_t)^* - (R_t - \bar{Q}_t)(R_t - \bar{Q}_t)^*}{(\bar{Q}_t - \bar{Q}_t)^* + \bar{Q}_t (\bar{Q}_t - \bar{Q}_t)^*} \tag{3.11}
\]

We can see that \( ss_t \) is proportional to the difference between the resolution and reliability for the \( t \)-th subsample. But this proportionality is not linear, because the denominator in Eq. 3.11 also depends on \( \bar{Q}_t \), which relates to the manner in which the forecaster sorts the \( K \) occasions into \( T \) subsamples.
Another measure of skill also has been found to be very useful. This measure is defined as follows:

\[ s_t = \frac{(RPS_t^c - RPS_t) - (R_t - \bar{R}_t)(R_t - \bar{R}_t)^*}{\bar{R} - R} \]  

(3.12)

The skill score \( s_t \) is linearly proportional to the difference between resolution and reliability for the \( t \)-th subsample. It has the property that the weighted average of \( s_t \) is the overall skill \( (s) \) with respect to the sample climatology (cf. Eq. 3.1 and Eq. 3.12). Thus,

\[ s = \frac{1}{K} \sum_{t=1}^{T} K_t s_t \] 

(3.13)

Therefore, by using Eq. 3.13, we can determine which subsamples make the largest contributions to the overall skill. The percentage contribution of each subsample to the overall skill \( (sp_t) \) will be calculated for all cases by using

\[ sp_t = \frac{K_t s_t}{\|s\|} \] 

(3.14)

Note that the absolute value of \( s \) has been used in the denominator of Eq. 3.14. Thus, for positive overall skill the percentage contribution from all subsamples sums up to 100%, whereas \( \sum_{t=1}^{T} sp_t = -100\% \) for negative overall skill. Therefore, positive values of \( sp_t, ss_t \), and \( s_t \) will always mean positive skill for the \( t \)-th subsample.
As pointed out by Brier (1950), the most useful forecasts are those which fall into the extreme classes of probability values. The skill score $s_t$ is a measure of this "usefulness". For example, one forecast in the first subsample having a value of $s_1$ of 70% contributes as much to the overall skill as 70 forecasts in the second subsample having a value of $s_2$ of 1%. Thus, the forecasts in the first subsample can be thought of as 70 times more "useful" than the forecasts in the second subsample. As will be shown later in Chapter 6, $s_t$ always tends to have large absolute values in the subsamples for which the $R_t$'s contain some extreme probability values (i.e., probabilities close to zero or one). From the above discussion, it is apparent that $s_t$ is a much more useful measure of skill than $ss_t$. We shall base our discussion of skill in Chapter 6 mainly on $s_t$ rather than on $ss_t$. 
CHAPTER 4

Some Theoretical Properties of the RPS in the Two-State Situation

In this chapter we shall investigate the properties of the RPS in the two-state situation from a different point of view. In addition, a beta model is developed to obtain a forecaster's predictive distribution in an idealized situation by using these theoretical properties of the RPS.

Section 4.1 demonstrates that the RPS can be considered to represent (approximately) the variance of accuracy. Section 4.2 describes some of the properties of the RPS by examining several special forecasts. Section 4.3 treats accuracy as a random variable and describes its distribution function. Section 4.4 presents a beta model
for a forecaster's predictive distribution, and Section 4.5 gives an example of an application of this model.

4.1 The RPS as the Variance of Accuracy

In the two-state situation the forecast is \((R_1, 1)\), which can be treated as a scalar \(R = R_1\). The vector representing the observed state \(Q = (D_1, 1)\) can be also reduced to \(D = D_1\) in the same manner. Now, \(R\) represents the probability assigned by the forecaster to the event "rain" and \(D = 1 (0)\) represents the occurrence (non-occurrence) of this event.

We can think of \(R\) and \(D\) as two random variables. Then, the random variable \(R - D\) is a natural measure of the attribute accuracy defined in Chapter 3. To describe the distribution or behavior of any random variable, we are usually interested in examining at least two parameters; namely, the mean and the variance of the variable. In this case, we can estimate these two statistics of the distribution of \(R - D\) by collecting \(K\) forecasts and calculating sample estimates of the mean and the variance.
An estimate of the mean of R-D, denoted by \( \hat{E}(R-D) \), can be calculated by using

\[ \hat{E}(R-D) = \left( \frac{1}{K} \right) \sum_{k=1}^{K} (R_k - D_k) \ldots \] (4.1)

The variance of R-D, \( \text{Var}(R-D) \), can be written as

\[ \text{Var}(R-D) = E((R-O)^2) - (E(R-D))^2 \ldots \] (4.2)

Thus, an estimate of \( \text{Var}(R-D) \) can be written as

\[ \hat{\text{Var}}(R-D) = \hat{E}((R-O)^2) - (\hat{E}(R-D))^2 \ldots \] (4.3)

where

\[ \hat{E}((R-O)^2) = \left( \frac{1}{K} \right) \sum_{k=1}^{K} (R_k - D_k)^2 \ldots \] (4.4)

Note that the right-hand side of Eq. 4.4 is just the RPS as defined in Eq. 3.3. By using the fact that

\[ E(R-O) = \left( \frac{1}{K} \right) \sum_{k=1}^{K} R_k = \left( \frac{1}{K} \right) \sum_{k=1}^{K} D_k = \bar{R} - \bar{D} \ldots \]

Eq. 4.3 now becomes

\[ \hat{\text{Var}}(R-D) = \text{RPS} - (\bar{R} - \bar{D})^2 \ldots \]

or

\[ \text{RPS} = \hat{\text{Var}}(R-D) + (\bar{R} - \bar{D})^2 \ldots \] (4.5)

For reasonably large sample sizes (large \( K \)), \( \bar{R} - \bar{D} \) generally is very close to zero, unless some systematic errors (over- or under-forecasting) exist. Besides, the forecaster can inflate or deflate his/her probabilities to make the forecasts reliable in-the-large after acquiring some experience in making this particular kind of forecast. Thus, Eq. 4.5 can be approximated as

\[ \text{RPS} \approx \hat{\text{Var}}(R-D) \ldots \] (4.6)
Considering R and D as individual random variables again, we have
\[ \text{Var}(R-D) = \text{Var} R - 2\times \text{Cov}(R,D) + \text{Var} D, \quad (4.7) \]
where \( \text{Cov}(R,D) \) denotes the covariance of R and D. In the next section, we will discuss \( \text{Var}(R-D) \) for several special kinds of forecasts which have the property that \( E(R) = E(D) \) (in terms of the parameter estimates, it can be rewritten as \( \bar{R} = \bar{D} \)). For these "unbiased" forecasts the RPS can be used as an estimate of \( \text{Var}(R-D) \) as indicated in Eq. 4.6.

4.2 \( \text{Var}(R-D) \) for Several Special Kinds of Forecasts

4.2.1 Forecasts and Observations Uncorrelated

For two-event situations, the reliability diagram is a very useful method of investigating one aspect of the forecaster's ability (Murphy and Winkler, 1977). It contains a plot of the relative frequency of occurrence of one of the events against the forecast probability of that event for specific probability values.
The horizontal line joining points A and B on the reliability diagram in Fig. 4.1 represents forecasts that are uncorrelated with the observed states. To see this result, note that the probability of occurrence of the event for a certain subsample \( \Pr\{O=1|R=R_t\} \) does not depend on the choice of \( R_t \) for points on this line. That is, \( \Pr\{O=1|R=R_t\} = \Pr\{O=1\} = E(O) \) for all \( t=1,\ldots,T \). For a very large collection of forecasts, the expression \( \Pr\{O=1|R=R_t\} = E(O) \) can be rewritten as \( \overline{O}_t = \overline{O} \) for all subsamples \( t=1,\ldots,T \). Var\( (R-O) \) for this kind of forecast is obtained by setting Cov\( (R,O) = 0 \) in Eq. 4.7. Thus,

\[
\text{Var}(R-O) = \text{Var} R + \text{Var} O. \tag{4.8}
\]

The skill of such forecasts is negative (to be discussed in the next chapter). In this case, the more the forecaster spreads out the forecast probabilities towards the "extreme" values of zero and one, the larger will be \( \text{Var}(R) \) and, therefore, the larger will be \( \text{Var} (R-O) \) (i.e., the worse will be the forecasts).
4.2.2 Sample Climatological Forecasts and the Definition of $s^*$

For the fixed forecast of $R = E(D)$, we have $\text{Cov}(R,D) = \text{Var} R = 0$, which is a special case of the type of forecasts discussed in Section 3.3.1 above. Eq. 4.8 then reduces to

$$\text{Var}(R-D) = \text{Var} D.$$  (4.9)

The point 0 in Fig. 4.1 represents this particular set of
forecasts. It is convenient to regard this set of sample climatological forecasts as reference forecasts and to use it as a standard of comparison for other forecasts. Let $RPS^C$ denote the RPS for the sample climatological forecasts. Because the RPS is negatively oriented (i.e., the larger the RPS, the worse the forecasts), the skill for all forecasts can be defined quite naturally as

$$s = \frac{(RPS^C - RPS)}{RPS^C}$$

(Murphy, 1977). It is now also possible to express the skill in terms of the parameters instead of the parameter estimates. Thus, the skill can be redefined as

$$s^* = \frac{(\text{Var } D - \text{Var}(R-D))}{\text{Var } D}$$

(4.10)

The sign of $s^*$ is determined by the numerator of Eq. 4.10, which can be expressed as follows by using Eq. 4.7:

$$\text{Var } D - \text{Var}(R-D) = 2\text{Cov}(R,D) - \text{Var } R$$

(4.11)

In order to have positive skill, Cov(R,D) must be greater than $(1/2)\text{Var}(R)$. Furthermore, the greater the correlation between R and D, the higher the skill score $s^*$. 


4.2.3 Perfectly Reliable Forecasts

The set of perfectly reliable forecasts satisfies the following condition:

\[ \Pr\{D=1|R=R_t\} = R_t, \text{ for } t=1,\ldots,T. \]  \hspace{1cm} (4.12)

For a very large collection of forecasts, the above condition can be written as

\[ \tilde{R}_t = R_t, \text{ for } t=1,\ldots,T. \]

Thus, this set of forecasts can be represented by the 45\(^\circ\) diagonal line in Fig. 4.1.

In practice, experienced forecasters should be able to formulate reliable probability forecasts in situations involving formal or informal feedback concerning their performance and very large samples of forecasts. In this regard, subjective PoP forecasts have been routinely formulated and disseminated to the general public by NWS forecasters for many years. Fig. 6.1 shows that the solid curve representing the reliability of a set of these forecasts is indeed quite close to the 45\(^\circ\) line.

\[ \text{Cov}(R,D) \text{ in the case of perfectly reliable forecasts is } \]

\[ \text{Cov}(R,D) = \text{Var } R \]  \hspace{1cm} (4.13)

(see Appendix 4.1 for the derivation). Substituting \text{Var } R for \text{Cov}(R,D) in Eq. 4.7, we have
\[ \text{Var}(R-D) = \text{Var} C - \text{Var} R. \]  
\[ (4.14) \]

Here, a larger variance of \( R \) implies a better score (recall that the RPS is an estimator of \( \text{Var}(R-D) \)), as opposed to the case of \( \text{Cov}(R,D)=0 \) (cf. Eq. 4.8).

Expressing \( \text{Var}(R-D) \) in terms of \( s^* \) by using Eq. 4.10, Eq. 4.14 becomes

\[ \text{Var} R = s^* \text{Var} D \]  
\[ (4.15) \]

for perfectly reliable forecasts. From this equation, we can see more clearly that skill \( s^* \) is proportional to the degree of the dispersion of forecast probabilities under the assumption of perfect reliability.

4.2.4 Perfect Forecasts

Perfect forecasts are forecasts that make \( R^*_k = D^*_k \) for all \( k \) \((k=1,\ldots,K)\). It is a special case of perfectly reliable forecasts. Since \( R \) and \( D \) become the same variable, Eq. 4.13 becomes

\[ \text{Cov}(R,D) = \text{Var} R = \text{Var} D. \]  
\[ (4.16) \]

Thus,

\[ \text{Var}(R-D) = \text{Var} R - 2 \cdot \text{Cov}(R,D) + \text{Var} D = 0 \]

\( \text{Cov}(R,D) \) compensates for the variance of \( R \) and the variance of \( D \) completely in this case.
4.3 The Probability Distribution of $R_0$

Because the observation $D$ takes on only two values (0 and 1), the shape of the distribution function of $R_0$ depends largely on the distribution of $R$. The value of $D$ determines the sign of the variable $R_0$, since the value of $R$ is always between 0 and 1. Thus, we can think of the distribution of $R_0$ as a combination of two branches; namely, a positive branch and a negative branch. The positive branch is the distribution of the variable $R_0$ when $D=0$, and this branch can be denoted as the distribution of the random variable $R_1D=0$. The negative branch is the distribution of the random variable $R_{-1}D=1$. The means and the variances of these two distributions will be calculated in the following subsections.

4.3.1 The Means of $R_1D=0$ and $R_{-1}D=1$

The means of $R_1D=0$ and $R_{-1}D=1$, under the assumption of unbiasedness (i.e., $E(R) = E(D)$), can be expressed as follows:

\[
E(R_{1D=0}) = \frac{(\text{Var } D - \text{Cov}(R,D))}{(1-E(D))} \quad (4.17)
\]

\[
E(R_{-1D=1}) = -\frac{(\text{Var } D - \text{Cov}(R,D))}{E(D)} \quad (4.18)
\]

(see Appendices 4.3 and 4.4 for the derivations). For the
case of \( \text{Cov}(R,D) = 0 \) (for example, the sample climatological forecasts),

\[
E(R|D=0) = \text{Var } D / (1-E(D)) = E(C)
\]

and

\[
E(R-1|D=1) = - \text{Var } D / E(D) = -(1-E(D))
\]

(to obtain these two equations, we have used the equation

\[
\text{Var}(D) = E(D)(1-E(D)),
\]

which was proved in Appendix 4.2). These quantities are shown by the points \( O_0 \) and \( O_1 \) in Fig. 4.2.

![Fig. 4.2. The domain of R=0.](image)

As the correlation between \( R \) and \( D \) increases from 0 to 1, the two points \( O_0 \) and \( O_1 \) approach the point \( O \), which has a value of zero, as implied by Eqs. 4.17 and 4.18. This fact is shown by the two arrows in Fig. 4.2.

For perfectly reliable forecasts, \( \text{Cov}(R,D) = \text{Var } R = s^* \text{Var } D \), Eqs. 4.17 and 4.18 reduce to

\[
E(R|D=0) = (1-s^*)E(C)
\]

\[
E(R-1|D=1) = -(1-s^*)(1-E(D)).
\]

That is, the distances of the points \( O_0 \) and \( O_1 \) in Fig. 4.2 from the point \( O \) are linearly proportional to \( s^* \), the
measure of skill.

4.3.2 The Variances of R10=0 and R-110=1

The general expressions for the variances of R10=0 and R10=1 are very complicated. As a result, we shall consider these variances only in the special case of perfectly reliable forecasts. With some manipulation (see Appendix 4.5), it is possible to show that the variances of R10=0 and R-110=1 in this case satisfy the following relationship:

\[ s^*(1-s^*)E(0)(1-E(0)) = (1-E(0))Var(R10=0) + E(0)Var(R-110=1) \]

Solving this equation for \( s^* = 0 \) and \( s^* = 1 \), we have \( Var(R10=0) = Var(R-110=1) = 0 \) for both cases (since the variance is always positive). Fig. 4.3 shows the distribution function of R=0 for these two cases.

The right branch of R=0 for \( s^* = 0 \) in Fig. 4.3 represents the distribution function of R10=0. The left branch for \( s^* = 0 \) is the distribution function for R-110=1. In the case of perfect forecasts (\( s^* = 1 \)), the distribution of R=0 is combined into a single spike in the center (R=0=0).
Fig. 4.3. The distributions of R-D for $s^* = 0$ and $s^* = 1$ (assuming perfectly reliable forecasts).

If the type of distribution function of the forecast probability $R$ is specified, $\text{Var}(R|D=0)$ and $\text{Var}(R-1|D=1)$ can be calculated. Thus, we can conclude that the RPS (an estimator of $\text{Var}(R-D)$) is especially useful in that it can be used to approximate the distribution of R-D provided that the forecasts are perfectly reliable (which is not an unreasonable assumption in many cases) and that the form of the distribution function of $R$ is specified. The following section describes how beta distributions can be used to approximate the distribution of $R$ and how the distribution of $R-D$ can also be approximated by using this beta model.
4.4 A Beta Model for R=0

The beta\((a, b)\) distribution function takes the form

\[
  f(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1},
\]

where \(0<x<1, a>0, b>0,\) and \(B(a, b) = \int_0^1 y^{a-1} (1-y)^{b-1} dy.\) There are two advantages of using the beta distribution to approximate the distribution of the forecast probability \(R\). First, the domain of the beta distribution is limited to the interval \((0,1)\), which is exactly the same interval as the domain of \(R\). The second advantage is that this distribution allows two modes when the variance is large and only one mode when the variance is small. This property is desirable since a double maxima of the density function of \(R\) occurs for skill \(s^*\) close to one (perfect forecasts) and a single maximum occurs for \(s^*\) small.

For perfectly reliable forecasts, \(R\) can be approximated by a beta distribution in which \(a = E(D)(1-s^g)/s^*\) and \(b = (1-E(D))(1-s^g)/s^*\) (see Appendix 4.6 for the derivation of the parameter estimates). \(R\|D=0\) then also has a beta distribution with parameters \((a, b+1)\) and \(R\|D=1\) has a beta distribution with parameters \((a+1, b)\) (see Appendices 4.7 and 4.8).
The figures in Appendix 4.9 show the distributions of \( R, \, R_{ID}=0, \) and \( R_{ID}=1 \) for different values of \( s^* \) when \( E(D) = 0.3 \) (recall that \( s^* \) is determined by \( \text{Var}(R-D) \)). It is interesting to note that the means of \( R_{ID}=0 \) and \( R_{ID}=1 \) approach zero linearly as \( s^* \) becomes larger. Also, as \( s^* \) varies from 0 to 1, the shape of \( R \) changes from one maximum to two maxima, whereas the shape of \( R-D \) varies from two maxima to one maximum. The following section gives an example of the application of this idealized model.

4.5 An Example

Both the MOS and SUB forecasts for measurable rainfall (\( \geq 0.01 \) inches) in this experiment are fairly close to perfectly reliable forecasts (see Fig. 6.1). The average frequency of rain (\( \bar{D} \)) is 0.154. The skill scores (\( s \)) for the MOS and SUB forecasts are 27% and 32%, respectively (obtained by calculating the RPS).

The dashed lines in Fig. 4.4 and Fig. 4.5 show the beta distributions of \( R \) given the mean \( \bar{D} \) and the variance \( s^*\text{Var}(D) \) for the MOS and SUB forecasts. The solid lines on these figures show the empirical distributions obtained
by plotting $K_t$ against $R_t$. It can be seen that for small values of $R_t$, the beta distribution (based on an assumption of perfectly reliable forecasts) does not fit the empirical observations very well. But, significant characteristics of the distribution of $R-O$ can still be modeled quite well by using the beta distributions as shown in Fig. 4.6 and Fig. 4.7. The solid line in Fig. 4.6 is the observed empirical distribution of $R-O$. It represents the number of cases for the observation of certain values of $R-O$ (for MOS). The right branch of the dashed line in Fig. 4.6 represents the beta$(a,b+1)$ distribution, where $a$ and $b$ are the parameters used in the beta distribution of $R$. The left branch of the dashed line is the beta$(a+1,b)$ distribution. Fig. 4.7 shows the same graphs for the SUB forecasts. Recalling that only the skills (or the $\sqrt{\text{Var}(R-O)}$) and $\bar{D}$ are used to calculate these beta distributions, we can see that this model is a convenient way to approximate the forecaster's predictive distribution (i.e., frequency distribution of use of probability values).
Fig. 4.4. The distributions of $R$ for the POS forecasts (solid curve) and the beta model (dashed curve) for the threshold of 0.01 inches.
Fig. 4.5. The distributions for R for the SUB forecasts (solid curve) and the beta model (dashed curve) for the threshold of 0.01 inches.
Fig. 4.6. The distributions of R-D for the MOS forecasts (solid curve) and the beta model (dashed curve) for the threshold of 0.01 inches.
Fig. 4.7. The distributions for R-D for the SUB forecasts (solid curve) and the beta model (dashed curve) for the threshold of 0.01 inches.
CHAPTER 5

Methods of Evaluation

The problem of evaluating forecasts in a scientific and generally acceptable manner has proven to be almost as difficult as the problem of forecasting the weather itself (Murphy 1977). Even for the two-state situation, evaluation methods frequently may not lead to definite conclusions concerning a forecaster's (or forecast system's) ability. In the experiment of concern here, the problem is more complicated because each forecast is expressed in the form of a five-component vector (the final "1" in R (see Section 2.2.2.1) is not counted as a component). Each component is a probabilistic forecast for one of the five (cumulative) states. Even if we have some reasonable procedures to evaluate forecasts in this five-state situation, the results may be difficult to
interpret. For this reason, scalar evaluation and simplified three-state vector evaluation methods were used here. Scalar methods refer to the evaluation of each one of the 5 components in the cumulative forecast \( R \) separately (i.e., each component is treated as a forecast in a two-state situation). This scalar approach has also been used to evaluate the forecasts as noncumulative probabilistic forecasts \( C \) as well. Vector methods refer to the evaluation of three-state forecasts by selecting two components of \( R \) and treating these two components (together with a "1" for the last cumulative state) as a single forecast.

As stated in Chapter 3, the RPS is our basic evaluation measure. In this chapter, Sections 5.1 and 5.2 describe scalar and vector evaluation methods, respectively.

5.1 Scalar Methods

Each forecast is in the form \( R = (R_1, R_2, R_3, R_4, R_5, 1) \) in this experiment, where \( R_n \) is the forecast of rainfall amount greater than some threshold value \( x_n \) inches in a 12-hour period \( (x_1 = 2.00, x_2 = 1.00, x_3 = 0.50, x_4 = 0.25, \)
Scalar methods treat \( R = (R_n, 1) \) as in a two-state situation \((n=1, \ldots, 5)\). The following subsections describe geometric interpretations of the attributes of forecasts on the reliability diagram and a significance test of reliability for each subsample.

### 5.1.1 Geometrical Interpretation of the Attributes of Probability Forecasts on a Reliability Diagram

Let \( RE_t \) and \( RS_t \) denote the reliability and the resolution, respectively, for the \( t \)-th subsample. Then,

\[
RE_t = (R_t - \bar{R}_t) (\bar{R}_t - \bar{\bar{R}}_t)^* \quad (5.1)
\]

and

\[
RS_t = (\bar{R}_t - \bar{\bar{R}}) (\bar{\bar{R}}_t - \bar{\bar{R}})^* \quad (5.2)
\]

(see Section 3.2.2). In the reliability diagram for the two-state situation in Fig. 5.1, the point \( \bar{R}_t \) is plotted against the corresponding forecast probability \( R_t \) for the \( t \)-th subsample. The vertical distance between the point \( (R_t) \) on the 45° perfect reliability line and the point \( \bar{R}_t \) is the square root of the reliability in Eq. 5.1. The vertical distance between the point \( (\bar{R}) \) on the horizontal no correlation line and the point \( \bar{R}_t \) is the square root of the resolution in Eq. 5.2. The sign of the skill score \( s_t \) is determined by the relative magnitudes of these distances (see Eq. 3.12). In the example in Fig. 5.1, \( RS_t \)
is larger than \( RE_t \). Thus, the skill \( s_t \) for this subsample is positive.

Perfectly reliable forecasts
\[ s = \frac{\text{Var}(R)}{\text{Var}(D)} \]

Covariance forecasts
\[ s = 0, \quad \text{Cov}(R, D) = \frac{\text{Var}(R)}{2} \]

No-correlation forecasts
\[ s = -\frac{\text{Var}(R)}{\text{Var}(D)}, \quad \text{Cov}(R, C) = 0 \]

Fig. 5.1. Reliability diagram and the geometrical representation of resolution, reliability, and skill in the two-state situation.

The dashed line in Fig. 5.1 joins the points \( \bar{D}_t \) for which the two distances - the square roots of the reliability and the resolution - are equal in magnitude for all subsamples. It can be referred to as the no-skill line, since any \( \bar{D}_t \) that falls on this line will lead to a zero value of \( s_t \) for that subsample. It is a useful reference line to judge the sign of \( s_t \). It separates the
reliability diagram in the vertical direction into two regions, one represents positive skill and the other represents negative skill. Any $\tilde{D}$ points that lie on the same side of this dashed line as the 45° line have positive skill. Any $\tilde{D}$ points that lie on the same side of this dashed line as the horizontal no-correlation line have negative skill.

The sign of the overall skill score (s) also can be roughly judged by comparing the average relative frequencies ($\tilde{D}$'s) with this dashed no-skill line. A set of forecasts that falls exactly on this line will have zero skill, since each subsample of the sample will have no skill. The covariance between R and D for this set of forecasts is \( \text{Var}(R)/2 \) from Eq. 4.11 (under the assumption of unbiasedness in-the-large). We can see from Fig. 5.1 that as the orientation of the reliability curve rotates from the horizontal line to the 45° perfect reliability line with point 0 (i.e., the overall relative frequency $\tilde{D}$) fixed, the correlation between R and D increases and the skill varies from negative to positive values. This type of comparison is useful in evaluating the results of the QPF experiment, because the probabilities for some events (such as heavy rainfall) tend to be overforecast to the degree that the overall skill becomes negative. Those forecasts are highly unreliable and the orientation of the
corresponding curves on the reliability diagram are far from the 45° perfect reliability line.

5.1.2 Test of Reliability

Due to sampling variability, the average frequency in the subset \( \bar{D}_t \) can still deviate from the 45° line even though the forecast in the subset \( R_t \) is reliable. This situation can arise more frequently when the sample size \( K_t \) is small. Therefore, a statistical test of significance concerning reliability, taking into account the effect the sample size, will be performed on the experimental results.

In reality, such a test is concerned with whether to accept or reject the null hypothesis \( H_0 \) that the subset forecast is reliable; that is,

\[
H_0 : \Pr\{D=1|R=R_t\} = R_t
\]

The so-called observed level of significance in this situation is the probability that we reject the hypothesis \( H_0 \) when we should have accepted it. Thus, the smaller the observed level of significance the stronger the evidence in support of the conclusion that the forecast is unreliable. An observed level of significance of less
than 5% is considered small enough for confidence in rejecting the null hypothesis (or in identifying a subsample of forecasts as unreliable). On the other hand, a large value of the observed level of significance does not guarantee that the forecast is unreliable. The observed level of significance is calculated for each subsample by using the method described below.

Let $z$ equal $\frac{Z}{\sqrt{\frac{K_t}{R_t}}}$, and let $Z$ denote the random variable that has the outcome of $z$. Then

$$H_0: \Pr\{D=1\mid R_t\} = R_t$$

and

$$H_1: \Pr\{D=1\mid R_t\} \neq R_t.$$ 

Thus, the observed level of significance ($E$) is

$$E = 2 \Pr_{H_0} (Z > z), \text{when } z > K_t R_t$$  \hspace{1cm} (5.3)

$$E = 2 \Pr_{H_0} (Z < z), \text{when } z < K_t R_t.$$  \hspace{1cm} (5.4)

The above expressions can be interpreted as the probability of the outcome being worse (further away from the 45° line) than the present outcome $z$ if the forecast is reliable. The multiplier 2 arises from the fact that the error should be counted for the other side as well (thus, this test is two-sided). The probabilities in the above equations are calculated by assuming that the cases in a subsample are independent of each other. Then, $Z$ will have a binomial distribution with parameters $(K_t, R_t)$ (conditional on the observed sample size $K_t$). Although
the assumption of independence may not hold in general, it is believed the lack of independence is not serious enough to affect the results appreciably. In this regard, forecasts of $R_t$ generally will not have been issued consecutively in the QPF experiment (except perhaps in the case of the 0% forecasts for which the test has not been performed). A normal distribution instead of a binomial distribution is used when the sample size satisfied the criterion $\min(K_tR_t,K_t(1-R_t))>5$ to save computing time.

5.2 Vector Methods

In general, a single forecast in the QPF experiment is represented by the vector $\mathbf{B}=(R_1,R_2,R_3,R_4,R_5,1)$. We now simplify the six-state situation into two different three-state situations and describe vector evaluation in terms of these two situations. They are $(R_4,R_5,1)$ and $(R_2,R_3,1)$. The situation $(R_4,R_5,1)$ corresponds to cumulative forecast of the three states involving rainfall amounts greater than 0.25 inches, rainfall amounts between 0.01 inches and 0.25 inches and no rainfall or trace (all in a 12-hour period). These states occur much more frequently than other combinations of three states involving heavier rainfall amounts. Thus, the average
relative frequency vector \( \mathbf{\tilde{Q}} \) for these states will not be as close to the extreme values of 0 or 1 as the other possible combinations of three of the 6 states. We shall call this three-state situation the 4-5-6 case. Evaluation of the \((R_2,R_5,1)\) forecasts will examine the forecasters' ability in heavy rainfall situations together with the more moderate rainfall situations. We shall call this three-state situation the 2-5-6 case. The following two subsections describe the geometrical framework for the three-state RPS and the geometrical interpretations of attributes of probability forecasts in this framework.

5.2.1 A Geometrical Framework for the Three-State RPS

In general, we would denote \( \mathbf{R} \) in a three-state situation by \((R_1,R_2,1)\), where \( R_1 \) is the forecast probability of heavy rainfall and \( R_2 \) is the cumulative forecast probability of heavy and moderate rainfall. For convenience here, we express \( \mathbf{R} \) simply as \((R_1,R_2)\), since the last scalar component in \( \mathbf{R} \) is always equal to 1. Similarly, we denote \( \mathbf{Q} \) as \((Q_1,Q_2)\). We can then express \( \mathbf{R} \) as a point in a two-dimensional cartesian coordinate system as in Fig. 5.2. Since \( R_2 \) must be greater than \( R_1 \) and since the values of \( R_1 \) and \( R_2 \) are both between 0 and
Fig. 5.2. Geometrical framework for the RPS in the three-state situation.
1, the domain of all possible forecast probability values (points) takes the shape of a right, isosceles triangle. The three possible states of occurrence (\(D\), (0,0) which represents no precipitation, (0,1) which represents moderate rainfall, and (1,1) which represents heavy rainfall, are the three vertices of the triangle as shown in Fig. 5.2.

Extreme values of the components of \(R\) are associated with the three sides of the triangle. Subsamples corresponding to the left side of the triangle are forecasts in which the probability of occurrence of the first state (heavy rainfall) is zero. Subsamples corresponding to the upper side of the triangle are forecasts in which the probability of occurrence of the third state (no precipitation) is zero. Subsamples corresponding to the longest side of the triangle are forecasts in which the probability of occurrence of the second state (moderate rainfall) is zero.

The RPS for an individual forecast is simply the square of the euclidean distance between \(R\) and \(D\) in this triangle (Murphy and Stael von Holstein, 1975). This result can be easily proved by using the definition of the RPS (see Eq. 3.2 for \(K=1\)). As an example, consider the forecast \(\mathcal{F} = (R_1, R_2) = (0.1, 0.4)\), which corresponds to \(\mathcal{L} =\)
(0.1, 0.3, 0.6). This forecast will have a better score when the state of no precipitation occurs, since the distance between R and (0, 0) is less than the distances between R and the vertices corresponding to the other two observations.

There are two points that warrant special attention in this triangle. One such point is the uniform forecast \( \pi = (1/3, 1/3, 1/3) \), which will have the same Brier score \( \frac{1}{3} \sum (r_{n} - d_{n})^{2} \) no matter what state occurs. In Fig. 5.2 the point \( (1/3, 2/3) \) corresponds to this uniform forecast. It can be easily seen that this point is a better forecast if the second state occurs when the RPS is used as the evaluation measure, since the distance is shorter between \( (1/3, 2/3) \) and \( (0, 1) \) than between \( (1/3, 2/3) \) and \( (1, 1) \) or \( (0, 0) \). From this example, we can see that the RPS is a more reasonable scoring rule than the Brier score when ordered events are of interest. The only point that gives the same RPS, no matter which state occurs, is the point \( (1/2, 1/2) \) in Fig. 5.2. The distance between \( (1/2, 1/2) \) and each vertex is \( \sqrt{2}/2 \). The point \( (1/2, 1/2) \) corresponds to the (noncumulative) forecast of \( \pi = (1/2, 0, 1/2) \). As will be shown in the next subsection this point corresponds to a two-state forecast \( R=0.5 \) for which the "uncertainty" is the greatest.
5.2.2 Geometrical Interpretation of Attributes of Probability Forecasts

5.2.2.1 Uncertainty Term

The uncertainty term in the three-term partition of the RPS in Section 3.2.2.2 can be rewritten as (see Appendix 5.1)

\[ \bar{q}(u-\bar{Q})^2 = 0.5 - (0.5 - \bar{Q})(0.5 - \bar{Q})^2, \]

(5.5)

where \( \bar{Q} = (0.5, 0.5) \). Thus, the uncertainty term is the constant 0.5 minus the square of the euclidean distance between points \((0.5, 0.5) \) and \( \bar{Q} \). It is a function of \( \bar{Q} \) only. Since \( \bar{Q} \) has the same domain as \( R \), we can plot isolines of the uncertainty term as a function of \( \bar{Q} \) using the same geometrical framework as in the previous subsection. Fig. 5.3 shows such a plot. For example, if the average frequency (\( \bar{Q} \)) has the same coordinate as point "A" in Fig. 5.3, this set of forecasts will have an uncertainty score of 0.35, since point "A" lies mid-way between isolines 0.3 and 0.4. It is obvious that \((0.5, 0.5) \) is the point of maximum uncertainty.
Uncertainty can also be interpreted as the RPS for sample climatology (see Eq. 5.4). Sample climatology is the reference procedure with which we will compare the experimental forecasts. We can see that the closer \( \bar{q} \) is to \((0.5, 0.5)\), the "more difficult" it is to improve our forecasts. That is, in such a situation, we will have to reduce a larger value of RPS to have perfect forecasts (or, equivalently, there is more uncertainty involved in
predicting which state will occur). The uncertainty term is indeed a measure of uncertainty.

5.2.2.2 Resolution, Reliability, and Skill

For a subsample forecast $R_t$, reliability and resolution are the squares of the distances shown in Fig. 5.4 (see Eqs. 5.1 and 5.2). The skill ($s_t$) for $R_t$ in Fig. 5.4 is positive, since the resolution distance is larger than the reliability distance. The dashed line in Fig. 5.4 joins all the possible $\bar{R}_t$ values that will make the resolution and reliability equal in magnitude for this subsample forecast $R_t$. Thus, this line can be referred to as the no-skill line for $R_t$. Any $\bar{R}_t$ corresponding to this $R_t$ that lies to the right (in this example) of the dashed line will make the subsample skill positive. Any $\bar{R}_t$ that lies to the left of the dashed line will make the subsample skill negative. Also, the distance between $\bar{R}_t$ and the dashed line ("a" in Fig. 5.4) is proportional to the skill $s_t$ (proved in Appendix 4.2). Thus, the further $\bar{R}_t$ is away from this no-skill line the larger (positively or negatively) is the skill value.
Fig. 5.4. Geometrical representation of resolution, reliability, and skill in the three-state situation.
CHAPTER 6

Experimental Results

The experimental results are presented in this chapter. Section 6.1 describes the results of scalar evaluation and Section 6.2 describes the results of vector evaluation. Scalar evaluation refers to the evaluation of each component of a cumulative forecast $\mathbf{A}$ separately, whereas vector evaluation refers to the evaluation of two or more components of $\mathbf{A}$ jointly. Because a large amount of data has been produced by using different evaluation procedures for different periods (day and night) and different stations, only selected results are reported here in the form of tables and figures.
6.1 Scalar Evaluation

This section is divided into several subsections. Each subsection will focus on a specific attribute or issue (e.g., day/night differences, station differences).

6.1.1 Sample Reliability

The average objective (MOS) and subjective (SUE) forecast probabilities and the observed relative frequencies for cumulative or noncumulative precipitation amount events in the first year of the quantitative precipitation forecasting experiment are presented in Table 6.1, together with the 20-year (long-term) climatological probabilities.

The four events representing precipitation amounts greater than 0.25 inches occurred only 4.7% of the time overall, whereas the event representing none or trace alone occurred almost 85% of the time. The overall correspondence between the MOS average probabilities and the observed relative frequencies is quite good. The SUE forecasts, however, show a tendency for average
Table 6.1. Climatological probabilities and average forecast probabilities for MOS and SUB. $\bar{D}$ denotes the observed relative frequencies (sample climatological probabilities) and CLI denotes the long-term climatological probabilities.

<table>
<thead>
<tr>
<th>EVENT</th>
<th>MOS</th>
<th>SUB</th>
<th>$\bar{D}$</th>
<th>CLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.00</td>
<td>.001</td>
<td>.008</td>
<td>.006</td>
<td>.003</td>
</tr>
<tr>
<td>21.00</td>
<td>.008</td>
<td>.019</td>
<td>.014</td>
<td>.011</td>
</tr>
<tr>
<td>20.50</td>
<td>.020</td>
<td>.042</td>
<td>.026</td>
<td>.026</td>
</tr>
<tr>
<td>20.25</td>
<td>.051</td>
<td>.084</td>
<td>.047</td>
<td>.045</td>
</tr>
<tr>
<td>20.01</td>
<td>.184</td>
<td>.180</td>
<td>.153</td>
<td>.140</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EVENT</th>
<th>MOS</th>
<th>SUB</th>
<th>$\bar{D}$</th>
<th>CLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.00</td>
<td>.001</td>
<td>.008</td>
<td>.006</td>
<td>.003</td>
</tr>
<tr>
<td>21.00</td>
<td>.008</td>
<td>.019</td>
<td>.014</td>
<td>.011</td>
</tr>
<tr>
<td>20.50</td>
<td>.020</td>
<td>.042</td>
<td>.026</td>
<td>.026</td>
</tr>
<tr>
<td>20.25</td>
<td>.051</td>
<td>.084</td>
<td>.047</td>
<td>.045</td>
</tr>
<tr>
<td>20.01</td>
<td>.184</td>
<td>.180</td>
<td>.153</td>
<td>.140</td>
</tr>
</tbody>
</table>

It is also of interest to note that the 20-year climatological probabilities are very close to the sample relative frequencies. Thus, the period of the experiment is similar to the 20-year historical period (with respect to these precipitation events).

6.1.2 Subsample Reliability

Figs. 6.1-6.5 contain the reliability diagrams for each cumulative event for all four stations and both periods combined. The three straight lines that meet at the point $R=\bar{D}$ (overall sample relative frequency) are the 45° perfect reliability line, the no-skill line, and the
Fig. 6.1. Reliability diagram for forecasts of precipitation ≥0.01 inches for two time periods (day/night) combined. The solid curve is for the SUB forecasts and the dashed curve is for the POS forecasts. The three straight lines are the perfect reliability line, the no-skill line, and the no-correlation line as in Fig. 5.1. $R$, $K$, $E$, $RE$, $RS$, $S$, and $SP$ denote the forecast probability, sample size, observed level of significance, reliability, resolution, skill, and skill contribution of a subsample, respectively.
### Forecast Probability Distribution

<table>
<thead>
<tr>
<th>Forecast Probability</th>
<th>Observed Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>90</td>
<td>100</td>
</tr>
</tbody>
</table>

**Fig. 6.2. Reliability diagram for forecasts of precipitation ≥0.25 inches.** See the legend in Fig. 6.1 for an explanation of the symbols, lines, and curves.

<table>
<thead>
<tr>
<th>Location</th>
<th>Precipitation (inches)</th>
<th>Observed</th>
<th>Expected</th>
<th>Rainfall Days</th>
<th>Accumulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 6.2.** Summary of precipitation data for various locations. The table provides the observed and expected precipitation amounts, along with the number of rainfall days and total accumulation for each location.
Fig. 6.3. Reliability diagram for forecasts of precipitation $\geq 0.50$ inches. See the legend in Fig. 6.1 for an explanation of the symbols, lines, and curves.
Fig. 6.4. Reliabiliy diagram for forecasts of precipitation ≥1.00 inches. See the legend in Fig. 6.1 for an explanation of the symbols, lines, and curves.
Fig. 6.5. Reliability diagram for forecasts of precipitation ≥2.00 inches. See the legend in Fig. 6.1 for an explanation of the symbols, lines, and curves.
horizontal no-correlation line discussed in the previous chapter. The solid curve is the reliability (or $\tilde{D}_k$) curve for the SUB forecasts and the dashed line is the reliability curve for the MOS forecasts. It can be seen that, for the event with a threshold value of 0.01 inches (see Fig. 6.1), both the MOS and SUB curves are quite close to the 45° line. However, for forecasts of larger rainfall events, the SUB curves “follow” the no-skill lines or fall between the no-skill line and the no-correlation line (e.g., see Fig. 6.5). This result indicates overforecasting for the heavier rainfall events by the SUB forecasts. On the other hand, the MOS curves remain quite reliable for all threshold values.

In judging the reliability of the forecasts in more detail, it is important to consider the sample sizes associated with the points (since sampling variability increases as sample size decreases). The subsample sizes ($K_w$) and the observed level of significance (E) in percent are shown in the tables below the reliability diagrams. The latter is included only when its value for a subsample is less than 5%. As stated in Chapter 4, those points associated with subsample forecasts with small $E$ values can be considered to represent unreliable forecasts. For forecasts of precipitation amounts greater than 0.01 inches ($R_x$, see Fig. 6.1), very few probabilities are
Identified as unreliable, although the sample sizes are quite large for most of the subsamples. For forecasts of amounts greater than 0.25 inches ($R_4$, see Fig. 6.2), most of the SUB forecasts are not reliable since 6 subsamples have observed levels of significance less than 5%. For heavier rainfall forecasts (see Figs. 6.3, 6.4, and 6.5), many subsamples in the SUB forecasts depart markedly from the 45° perfect reliability line, but the sample sizes associated with these subsamples are too small to conclude that the forecasts are unreliable.

It is interesting to note that the MOS forecasts are more conservative in the sense that the ranges of probability values used in the MOS forecasts are narrower than those used in the SUB forecasts for heavier rainfall events. However, the SUB forecasts were not very successful in increasing the variability of the forecast probabilities (i.e., in increasing the skill; see discussion related to Eq. 4.15), while maintaining reliable forecasts.
6.1.3 Subsample Skill

The reliability (REₜ), resolution (RSₜ), skill (sₜ), and skill contribution (spₜ) for each subsample are also shown below the reliability diagrams. REₜ and RSₜ are expressed in the units of $10^{-4}$ (percentages squared) and sₜ and spₜ are expressed in percentages. From the table in Fig. 6.1, it can be seen that the resolution for both MOS and SUB increases from the point R=0 (sample climatological forecasts) in both directions, whereas the reliability remains quite small and exhibits only some "random" variability. Thus, the skill shows the same trend as does the resolution. For reliable forecasts, the forecasts in the extreme subsamples (subsamples with forecast probability of zero or one) are more skillful since the magnitude of $sₜ$ is higher. The percentage contribution to the overall skill from each subsample (spₜ) exhibits a quite uniform distribution across probability values, although the subsample sizes are considerably smaller for forecasts with $Rₜ>50\%$ than for forecasts with $Rₜ<50\%$. For the SUB forecasts, 61% of the overall skill is contributed by the subsamples with $Rₜ$ greater than or equal to 60%, whereas only 42% of the overall skill is contributed by those subsamples for the MOS forecasts.
The relationships among $K_t$, $S_t$, and $SP_t$ for the MCS forecasts are shown in Fig. 6.6. We can see that $S_t$ is close to zero and $SP_t$ attains a minimum value for the 20% forecasts, a point that is close to the point for which $R=0$. (Recall that skill is measured relative to the sample climatological probabilities.) Also, it is clear that a larger subsample size does not guarantee a greater contribution to the overall skill. From the table in Fig. 6.2, it can be seen that there are 100 SUB forecasts of 30% probability that make a large negative (-155%) contribution to the overall skill. Similarly, the table in Fig. 6.4 for the SUB forecasts indicates that the subsample forecasts of greater than 60% (with only 5 cases) contribute -95% to the overall skill. These examples provide evidence that although the forecasts associated with "extreme" probabilities may be skillful, they are skillful only when they are reliable. The penalty for lack of reliability can be very high. Fig. 6.5 shows a similar result for the SUE forecasts for another threshold value.
Fig. 6.6. Distribution of $K_t$, $S_t$, and $SP_t$ for the MOS forecasts for the forecasts with a threshold of 0.01 inches. The dashed line is the curve for skill contribution ($SP_t$), the solid line with a peak on the right is the curve for the skill ($S_t$), and the other solid line is the curve for the subsample size ($K_t$).
6.1.4 Overall Skill and Day/Night Differences

Table 6.2 shows some summary results of overall skill. The columns denoted by "ALL" are the overall skill for the two periods (day/night) combined. For the cumulative event forecasts, we can see that skill decreases as the threshold value of the precipitation amount increases for both the MOS and SUB forecasts. The SUB forecasts were more skillful than the MOS forecasts only for the threshold of 0.01 inches. For events involving rainfall amounts greater than 1.00 inches, the SUB forecasts have negative overall skill.

<table>
<thead>
<tr>
<th>EVENT</th>
<th>MOS</th>
<th>SUB</th>
<th>MOS</th>
<th>SUB</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥2.00</td>
<td>1</td>
<td>-16</td>
<td>3</td>
<td>-20</td>
</tr>
<tr>
<td>≤1.00</td>
<td>7</td>
<td>-7</td>
<td>7</td>
<td>-34</td>
</tr>
<tr>
<td>≤0.50</td>
<td>9</td>
<td>1</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>≤0.25</td>
<td>16</td>
<td>2</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>≤0.01</td>
<td>26</td>
<td>32</td>
<td>26</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 6.2. The overall skill (in percent) for the MOS and SUB forecasts for different time periods.
Examination of the skill for the night and day periods indicates that the SUB forecasts were not very skillful for the nighttime period, whereas the MCS forecasts indicate little difference in skill between the daytime and the nighttime periods. As an illustration, Fig. 6.7 and Fig. 6.8 contain the reliability diagrams for forecasts of rainfall amounts greater than 0.25 inches for day and night, respectively. The difference between the two periods for the SUB forecasts is quite dramatic. The solid curve for daytime forecasts is not very far away from the 45° line, whereas the solid curve for nighttime forecasts falls almost completely in the negative skill region. For the MCS forecasts, the day and night curves show similar behavior. It should be noted that heavier rainfall occurs more frequently in the day than at night. For this threshold (forecasts of greater than 0.25 inches), the average relative frequency (D) is 3.4% for the nighttime and 5.9% for the daytime. It appears that the less frequent the event is the more severe is the degree of overforecasting (by the SUB forecasts). That is, the skill for the SUB forecasts is relatively poor for both the heavy rainfall amount forecasts and the nighttime forecasts. The skill of the noncumulative event forecasts shown in Table 6.2 is similar to that of the cumulative event forecasts.
Fig. 6.7. Reliability diagram for the daytime forecasts of precipitation ≥ 0.25 inches. See the legend in Fig. 6.1 for an explanation of the symbols, lines, and curves.
Fig. 6.8. Reliability diagram for the nighttime forecasts of precipitation ≥0.25 inches. See the legend in Fig. 6.1 for an explanation of the symbols, lines, and curves.
6.1.5 Three-Term Partition of the RPS

Table 6.3 shows the RPS, the terms in the partition of the RPS, and the skill scores for the MOS and SUB forecasts and also for the climatological forecasts (CLI) based on 20 years of historical data.

<table>
<thead>
<tr>
<th>EVENT</th>
<th>RPS</th>
<th>UNC</th>
<th>RE</th>
<th>RS</th>
<th>S</th>
<th>S_{CLI}</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\geq 2.00$ MOS</td>
<td>.006</td>
<td>.006</td>
<td>.0000</td>
<td>.0001</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>SUB</td>
<td>.007</td>
<td>.006</td>
<td>.0012</td>
<td>.0003</td>
<td>-.16</td>
<td>-.16</td>
</tr>
<tr>
<td>CLI</td>
<td>.006</td>
<td>.006</td>
<td>.0002</td>
<td>.0002</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>$\geq 1.00$ MOS</td>
<td>.013</td>
<td>.014</td>
<td>.0002</td>
<td>.0012</td>
<td>.07</td>
<td>.06</td>
</tr>
<tr>
<td>SUB</td>
<td>.015</td>
<td>.014</td>
<td>.0024</td>
<td>.0014</td>
<td>-.07</td>
<td>-.08</td>
</tr>
<tr>
<td>CLI</td>
<td>.014</td>
<td>.014</td>
<td>.0005</td>
<td>.0006</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>$\geq 0.50$ MOS</td>
<td>.023</td>
<td>.026</td>
<td>.0005</td>
<td>.0028</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>SUB</td>
<td>.026</td>
<td>.026</td>
<td>.0036</td>
<td>.0037</td>
<td>.01</td>
<td>-.01</td>
</tr>
<tr>
<td>CLI</td>
<td>.026</td>
<td>.026</td>
<td>.0010</td>
<td>.0013</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>$\geq 0.25$ MOS</td>
<td>.038</td>
<td>.045</td>
<td>.0007</td>
<td>.0079</td>
<td>.16</td>
<td>.15</td>
</tr>
<tr>
<td>SUB</td>
<td>.044</td>
<td>.045</td>
<td>.0056</td>
<td>.0078</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>CLI</td>
<td>.045</td>
<td>.045</td>
<td>.0018</td>
<td>.0020</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>$\geq 0.01$ MOS</td>
<td>.096</td>
<td>.130</td>
<td>.0032</td>
<td>.0375</td>
<td>.26</td>
<td>.28</td>
</tr>
<tr>
<td>SUB</td>
<td>.089</td>
<td>.130</td>
<td>.0020</td>
<td>.0431</td>
<td>.32</td>
<td>.33</td>
</tr>
<tr>
<td>CLI</td>
<td>.132</td>
<td>.130</td>
<td>.0103</td>
<td>.0083</td>
<td>-.02</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.3. The RPS, the terms in the partition of the RPS, and the skill scores for MOS, SUB, and CLI in the two-state (scalar) situation.

The uncertainty term is the largest term for all forecasts (i.e., all thresholds). It depends only on the relative frequency of the forecast event and is the same
for the MOS, SUB, and CLI forecasts. The resolution term for the SUB forecasts is larger (i.e., better) than the resolution terms for the MOS and CLI forecasts for all thresholds (except for the threshold of 0.25 inches). However, an examination of the magnitudes of the reliability terms reveals the opposite result. The reliability of the SUB forecasts is the poorest of the three types of forecasts except for the threshold of 0.01 inches. This result is obviously due to the overforecasting exhibited by the SUB forecasts for higher thresholds. We can see that if the forecasters could deflate their forecasts for heavy rainfall amounts (i.e., more closely approximate the 45° line in the reliability diagram), the SUB forecasts would be better than the MCS forecasts for every category since the skill would then depend only on the resolution term.

The reliability and resolution terms for the CLI forecasts tend to cancel each other for all thresholds. This fact makes the skill of the CLI forecasts with respect to the sample climatology very close to zero. Also, it is clear that there is little difference between using the sample climatology and using the long-term 20-year climatology as a standard of reference to calculate the skill scores in this experiment.
6.1.6 Forecasts for Four Stations

Table 6.4 shows the overall skill of the forecasts for the four stations. The skill of the SUE forecasts shows greater variability than the skill of the MCS forecasts. Skill seems to have a relatively close correspondence with the frequency of the forecast event (as in the case of day versus night). The driest station among the four stations is BRO and the wettest station is IAH. We can see from Table 6.4 that the skill for BRO forecasts is lower for both the MOS and SUE forecasts and for all thresholds than the skill for the IAH forecasts (except for the threshold of 2.00 inches).

6.2 Vector Evaluation

6.2.1 Three-Term Partition of the RPS and the Overall Skill

Table 6.5 shows the summary results from the vector evaluation. The three-term partition of the RPS and the overall skill is shown for different time periods and for both the 4-5-6 and 2-5-6 cases (see Section 5.2 for the
<table>
<thead>
<tr>
<th>EVENT</th>
<th>AUS</th>
<th>BRO</th>
<th>IAH</th>
<th>SAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥2.00</td>
<td>MOS -2</td>
<td>-1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>SUB -11</td>
<td>-3</td>
<td>-22</td>
<td>-22</td>
</tr>
<tr>
<td>≥1.00</td>
<td>MOS 4</td>
<td>3</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>SUB -22</td>
<td>-8</td>
<td>0</td>
<td>-6</td>
</tr>
<tr>
<td>≥0.50</td>
<td>MOS 7</td>
<td>3</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>SUB -9</td>
<td>-13</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>≥0.25</td>
<td>MOS 17</td>
<td>10</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>SUB 3</td>
<td>-6</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>≥0.01</td>
<td>MOS 29</td>
<td>15</td>
<td>25</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>SUB 33</td>
<td>22</td>
<td>32</td>
<td>37</td>
</tr>
</tbody>
</table>

Table 6.4. Overall skill (in percent) for four stations.

definition of these cases). Several conclusions can be drawn from this table: (1) The subjective forecasts have better resolution but poorer reliability than the MOS forecasts. (2) The magnitudes of $s$ and $s^{CL}$ are quite similar. (3) The skill of the subjective forecasts is better for the daytime than the nighttime forecasts, whereas the skill of the MOS forecasts is similar for the two time periods. (4) The uncertainty term makes the largest contribution to the magnitude of the RPS. These conclusions are similar to those deduced from the scalar evaluation.
<table>
<thead>
<tr>
<th></th>
<th>RPS</th>
<th>UNC</th>
<th>RE</th>
<th>RS</th>
<th>S</th>
<th>S_{CLI}</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-5-6</td>
<td>ALL MOS</td>
<td>.134</td>
<td>.175</td>
<td>.010</td>
<td>.052</td>
<td>.24</td>
</tr>
<tr>
<td>CASE</td>
<td>SUB</td>
<td>.133</td>
<td>.175</td>
<td>.016</td>
<td>.058</td>
<td>.24</td>
</tr>
<tr>
<td></td>
<td>CLI</td>
<td>.177</td>
<td>.175</td>
<td>.012</td>
<td>.010</td>
<td>-.01</td>
</tr>
<tr>
<td></td>
<td>NI. MOS</td>
<td>.116</td>
<td>.151</td>
<td>.015</td>
<td>.051</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>.127</td>
<td>.151</td>
<td>.028</td>
<td>.052</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>CLI</td>
<td>.154</td>
<td>.151</td>
<td>.013</td>
<td>.009</td>
<td>-.02</td>
</tr>
<tr>
<td>DAY</td>
<td>MOS</td>
<td>.148</td>
<td>.194</td>
<td>.014</td>
<td>.060</td>
<td>.24</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>.138</td>
<td>.194</td>
<td>.020</td>
<td>.077</td>
<td>.25</td>
</tr>
<tr>
<td></td>
<td>CLI</td>
<td>.196</td>
<td>.194</td>
<td>.012</td>
<td>.010</td>
<td>-.01</td>
</tr>
<tr>
<td>2-5-6</td>
<td>ALL MOS</td>
<td>.109</td>
<td>.144</td>
<td>.006</td>
<td>.042</td>
<td>.25</td>
</tr>
<tr>
<td>CASE</td>
<td>SUB</td>
<td>.104</td>
<td>.144</td>
<td>.009</td>
<td>.049</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>CLI</td>
<td>.148</td>
<td>.144</td>
<td>.013</td>
<td>.009</td>
<td>-.03</td>
</tr>
<tr>
<td></td>
<td>NI. MOS</td>
<td>.096</td>
<td>.128</td>
<td>.011</td>
<td>.042</td>
<td>.25</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>.098</td>
<td>.128</td>
<td>.013</td>
<td>.043</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>CLI</td>
<td>.130</td>
<td>.128</td>
<td>.011</td>
<td>.008</td>
<td>-.02</td>
</tr>
<tr>
<td>DAY</td>
<td>MOS</td>
<td>.120</td>
<td>.158</td>
<td>.008</td>
<td>.046</td>
<td>.24</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>.109</td>
<td>.158</td>
<td>.013</td>
<td>.061</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td>CLI</td>
<td>.163</td>
<td>.158</td>
<td>.014</td>
<td>.009</td>
<td>-.03</td>
</tr>
</tbody>
</table>

Table 6.5. The RPS, the terms in the partition of the RPS, and the skill scores for the MOS, SUB, and CLI forecasts in the simplified three-state situations.
6.2.2 Distribution of the Subsample Size and the Subsample Skill Contribution

6.2.2.1 MOS Forecasts

Fig. 6.9 shows the distribution of the subsample size ($K_t$) and skill contribution ($sp_t$) for the MOS forecasts for all four stations and two periods combined in the 4-5-6 case. The numbers in the figure are subsample skill contributions to the overall skill (in percent). They are placed in this space according to the corresponding forecast probabilities ($\pi_t$) of the subsample. A blank denotes the skill contribution is near zero (between -0.5% and 0.5%) for that subsample. The dashed curves are the contour lines of equal subsample size. They are not drawn at constant (or fixed) intervals, since most of the forecasts are concentrated near the point of average relative frequency $\bar{\pi}$ which is denoted by an asterisk ("*") in Fig. 6.9. Examination of the distribution of subsample sizes (isolines) reveals that the MOS forecasts tend to avoid using extreme probability values. Recall that the extreme values of $\pi_t$ are associated with the three sides of the triangle (see Section 5.2.1).
Fig. 6.9. The distribution of \( K_t \) and \( s_p_t \) for the MOS forecasts in the 4-5-6 case. The asterisk "*" denotes the point \( D \).
The skill contribution \( sp_t \) is evenly distributed and it is quite remarkable that none of the subsamples has a significant negative contribution to the overall skill. Again, we can see that positive skill is associated mostly with forecasts in subsamples that are further away from the sample relative frequency point \( \bar{Q} \). In fact, the contributions from subsamples near \( \bar{Q} \) are close to zero. Fig. 6.10 shows the distribution of \( k_t \) and \( sp_t \) for the MCS forecasts for the 2-5-6 case, and the results are similar in this case.

6.2.2.2 SUB Forecasts

The distribution of subsample size \( k_t \) for the SUB forecasts (not shown) appears to be less systematic than that for the MOS forecasts. More forecasts in the 4-5-6 case are scattered near the longest side of the triangle in \( R \) space for the SUB forecasts than for the MCS forecasts. The skill contribution \( sp_t \) is also not evenly distributed. This vector presentation of the results helps to locate systematic errors made by the SUB forecasts. An example is shown in Fig. 6.11.
Fig. 6.10. The distribution of $K_2$ and $SP_2$ for the MOS forecasts in the 2-5-6 case. The asterisk "*" denotes the point $Q$. 
Fig. 6.11. The spectra for the nighttime SUB forecasts in the 4-5-6 case. The enclosed area is associated with bimodal forecasts.
This figure presents the skill contribution for the nighttime SU8 forecasts. The area enclosed by the straight lines is associated with bimodal forecasts. That is, a forecast in this area assigns a probability to the second state that is smaller than the probabilities of both the first and the third state. An example of an E_t of this kind is shown in Fig. 6.12. We can conclude that this forecast is not reasonable unless the forecaster really thinks that either no rain (the third state) will occur or that heavy rain (the first state) will occur.

Examination of Fig. 6.11 reveals that such bimodal forecasts generally make negative contributions to the overall skill score. The problem may have been caused in part by a lack of understanding by the forecasters of the nature of the forecast probability distribution. Because the forecast is issued in a cumulative form, it might not appear to the forecaster that a (0.3,0.3) forecast, as in the example, leads to the distribution pictured in Fig. 6.12. It might also be that the forecasters were "trying too hard" to assign high probabilities to heavy precipitation.
Fig. 6.12. The probability function for a bimodal forecast $R=(0.3,0.3)$ ($\bar{c}=(0.3,0,0.7)$).

6.2.3 Subsample Reliability

Fig. 6.13 shows $R_{1k} - \bar{c}_{kt}$ (recall that $(R_t - \bar{c}_t) (R_t - \bar{c}_t)^*$ is the subsample reliability) for the SUB forecasts for all four stations and two periods combined in the 4-5-6 case. Fig. 6.14 shows the $R_{2k} - \bar{c}_{kt}$ for the same case. Subsamples whose sample size is less than 5 have not been included in the figures. By ignoring these points, we can draw contour lines for these quantities without having to face the difficulties caused by small subsample sizes. The quantities $R_{1k} - \bar{c}_{1k}$ and $R_{2k} - \bar{c}_{2k}$ can be thought of as
Fig. 6.13. $R_{it} - \bar{O}_{it}$ for the SUB forecasts in the 4-5-6 case.
Fig. 6.14. $R_{zt} - \bar{D}_{zt}$ for the SUB forecasts in the 4-5-6 case.
indicators of the degree of overforecasting. The hatched area represents the region with negative $R_{it} - \bar{O}_{it}$ values ($i=1,2$) (i.e., underforecasting).

It was found in the scalar evaluation that the SUB forecasts for the first state (precipitation amount greater than 0.25 inches) were not reliable and often assigned too high probabilities to this event. Fig. 6.13 shows that this deviation from perfectly reliable forecasts is most severe in or near the area of bimodal forecasts. Somewhat surprisingly, there are several subsamples that underforecast for the first state.

Fig. 6.14 shows the degree of overforecasting /underforecasting for forecasts of precipitation amounts greater than 0.01 inches. Although the scalar evaluation indicates that they are quite reliable (close to the 45° line), a vector evaluation indicates that $R_{2t} - \bar{O}_{2t}$ is still quite large and reveals some systematic patterns. The area of underforecasting (hatched area) seems to be as large as the area of overforecasting. As for the MCS forecasts, the pattern of reliability (not shown) is more random and the magnitude of the reliability "error" is smaller.
CHAPTER 7

Conclusion

The properties of the RPS have been investigated and several methods of evaluating probabilistic forecasts for ordinal events have been developed in this study. Moreover, the methods have been applied on the results of a QPF experiment undertaken by the forecasters at the San Antonio WSFO.

It has been shown that the distribution of accuracy ($R-Q$) can be approximated very well by using a beta model for samples of forecasts that are close to being perfectly reliable. This beta model requires only the RPS and the sample relative frequency ($\bar{Q}$) as its input information. The RPS is also very useful since the partition of the RPS provides measures of reliability, resolution, and skill.
A skill score for subsample forecasts has been devised. This score has the property of being linearly proportional to the difference between subsample resolution and reliability. As a result, it makes the calculation of "the percentage contribution of subsample skill to overall skill" possible. This method is useful in finding systematic errors made by forecasters if such errors exist.

In scalar evaluation, the inclusion of the no-skill line and the no-correlation line on the reliability diagram is helpful in representing the skill, reliability, and resolution geometrically in two-state situations. Geometrical interpretations of attributes of forecasts can also be accomplished in three-state situations by using vector evaluation procedures. It has been found that the uncertainty inherent in three-state situations is a maximum for sample relative frequencies equal to (0.5, 0, 0.5) (expressed noncumulatively).

The experimental results show that the skill of the SUB forecasts is generally less than the skill of the MCS forecasts except for forecasts of measurable precipitation (i.e., for PoP forecasts). This result is due primarily to the forecaster's tendency to overforecast for categories associated with larger precipitation amounts.
This tendency is most pronounced for the nighttime forecasts or the forecasts for drier stations. The MOS forecasts, however, are quite reliable for both periods (day/night). The vector evaluation approach has shown that the degree of overforecasting is quite high for bimodal forecasts and that the skill contribution from the bimodal forecasts is negative in many cases.

It should be kept in mind that the forecasters at San Antonio WSFO had no prior experience in making probabilistic QPFs and that they did not receive any formal feedback concerning their performance during the first year of the experiment. Feedback (e.g., feedback regarding the tendency to overforecast for larger precipitation amounts) was provided in January 1982 and the experiment ended in June 1982. Thus, a comparison of the experimental results for the 1981 and 1982 wet seasons (February - June) will be of considerable interest. With regard to other future analyses, the methodology developed in this study can be extended to the evaluation of non-cumulative forecasts using the Brier score, which is a suitable measure for situations involving non-ordinal variable forecasts. Also, it would be interesting to compare the results of these two methods of evaluating the QPF forecasts.


Murphy, A. H., 1970: The ranked probability score and the probability score: a comparison. J. Appl. Meteor., 9, 917-924.


APPENDICES

4.1. The proof that \( \text{Cov}(R, D) = \text{Var}(R) \) for perfectly reliable forecasts.

\[
E(RO|IR=IR_t) = R_t \text{Pr}\{O=1|R=IR_R_t\}.
\]

For perfectly reliable forecasts, \( \text{Pr}\{O=1|R=IR_R_t\} = R_t \).

Thus,

\[
E(RO|IR=IR_t) = R_t^2,
\]

\[
\text{Cov}(R, D) = E(RO) - E(R)E(D),
\]

\[
= E(E(RO|R)) - E(R)E(D).
\]

Using Eq. A1 and the equation \( E(R) = E(D) \) for unbiased forecasts, we have

\[
\text{Cov}(R, D) = E(R^2) - (E(R))^2,
\]

\[
= \text{Var}(R).
\]

4.2. The proof that \( \text{Var}(D) = E(D)(1-E(D)) \).

\[
\text{Var}(D) = E(D^2) - (E(D))^2,
\]

\[
= E(D) - (E(D))^2,
\]

\[
= E(D)(1-E(D)).
\]

4.3. The proof that \( E(R-D) = (\text{Var}(D) - \text{Cov}(R, D))/E(D) \) for unbiased forecasts \( (E(R) = E(D)) \).

\[
E(RD) = E(E(RDID)) = E(D)E(RDID=1) + (1-E(D))E(RDID=0),
\]

\[
= E(D)E(RID=1) + (1-E(D)*0,
\]

\[
= E(D)(E(R-ID=1) + 1).
\]
\[ \text{Cov}(R, D) = \text{E}(RD) - \text{E}(R) \text{E}(D), \]
\[ = \text{E}(D) \text{E}(R-D=1) + \text{E}(D) - \text{E}(R) \text{E}(C). \]

Substituting \( \text{E}(R) = \text{E}(D) \) into the above equation:
\[ \text{E}(R-D=1) = -(\text{E}(D)(1-\text{E}(D)) - \text{Cov}(R, D)) / \text{E}(D), \]
\[ = -(\text{Var}(D) - \text{Cov}(R, D)) / \text{E}(D). \]

(A2)

4.4. The proof that \( "\text{E}(R-D=0) = (\text{Var}(D) - \text{Cov}(R, D)) / (1-\text{E}(D)) \) for unbiased forecasts \( (\text{E}(R) = \text{E}(D))." \)
\[ \text{Cov}(R, D) = \text{E}((R-\text{E}(R))(D-\text{E}(D))), \]
\[ = \text{E}((R-\text{E}(D))(D-\text{E}(D))), \]
\[ = \text{E}(D) \text{E}((R-\text{E}(D))(D-\text{E}(D))(D=1)) \]
\[ + (1-\text{E}(D)) \text{E}((R-\text{E}(D))(D-\text{E}(D))(D=0)), \]
\[ = \text{E}(D)(1-\text{E}(D))(\text{E}(R-\text{E}(D))(D=1) - \text{E}(R-\text{E}(D))(D=0)), \]
\[ = \text{E}(D)(1-\text{E}(D))(\text{E}(R1D=1) - \text{E}(R1D=0)), \]
\[ = \text{E}(D)(1-\text{E}(D))(\text{E}(R-1D=1) - \text{E}(R1D=0) + 1). \]

Substituting Eq. A2 for \( \text{E}(R-1D=1) \), we have:
\[ \text{E}(R1D=0) = (\text{Var}(D) - \text{Cov}(R, D)) / (1-\text{E}(D)). \]

4.5. The proof that \( "s^*(1-s^*) \text{E}(D)(1-\text{E}(D)) = (1-\text{E}(D)) \text{Var}(R1D=0) + \text{E}(D) \text{Var}(R-1D=1) \) for unbiased forecasts \( (\text{E}(R) = \text{E}(D))." \)

From Eq. 4.10,
\[ (1-s^*) \text{Var}(D) = \text{Var}(R, D), \]
\[ = \text{E}(\text{E}((R-D-\text{E}(R-D))^2)(D))). \]
Since \( E(R-D) - E(D) \equiv 0 \),

\[
(1-s^2) \text{Var}(D) = E(E((R-D)^2 | D)) = (1-E(D))E((R-D)^2 | D=0) + E(D)E((R-D)^2 | D=1),
\]

\[
= (1-E(D))E(R^2 | D=0) + E(D)E((R-1)^2 | D=1).
\]

(A3)

\[
E(R^2 | D=0) = \text{Var}(R|D=0) + (E(R|D=0))^2.
\]

Substituting \( E(R|D=0) = (1-s^*)E(D) \) (Eq. 4.19) into the above equation, we have:

\[
E(R^2 | D=0) = \text{Var}(R|D=0) + (1-s^*)^2 (E(D))^2.
\]

For the same reason,

\[
E((R-1)^2 | D=1) = \text{Var}(R-1|D=1) + (1-s^*)^2 (1-E(D))^2.
\]

Substituting the above two equations into Eq. A3,

\[
(1-s^*) \text{Var}(D) = (1-E(D))(\text{Var}(R|D=0) + (1-s^*)^2 (E(D))^2) + E(D)(\text{Var}(R-1|D=1) + (1-s^*)^2 (1-E(D))^2),
\]

and rearranging the above equation:

\[
(1-s^*) \text{Var}(D) = (1-E(D))(E(D))^2 (1-s^*)^2 - E(D)(1-E(D))^2 (1-s^*)^2
\]

\[
= (1-E(D)) \text{Var}(R|D=0) + E(D) \text{Var}(R-1|D=1).
\]

Since \( \text{Var}(D) = E(D)(1-E(D)) \) (see Appendix 4.2), the above equation reduces to the following form:

\[
s^*(1-s^*)E(D)(1-E(D)) = (1-E(D)) \text{Var}(R|D=0) + E(D) \text{Var}(R-1|D=1).
\]

4.6. The proof of the statement that "for perfectly reliable forecasts, the beta(a,b) distribution has a mean of E(D) and a variance of \( s^* \text{Var}(D) \), where \( a = E(C)(1-s^*)/s^* \), \( b = (1-E(D))(1-s^*)/s^* \)."
The mean of a beta(a, b) distribution is:

\[
a = \frac{E(D)(1-s^*)/s^*}{a+b} = \frac{E(D)(1-E(D))}{a+b} + \frac{(1-E(D))(1-s^*)/s^*}{a+b}
\]

which is the mean of R, E(R). The variance of a beta(a, b) distribution is:

\[
ab = \frac{E(D)(1-E(D))(1-s^*)^2/s^2}{(a+b+1)(a+b)} = \frac{s^2}{(1-s^*)/(s^*+1)(1-s^*)/s^*}
\]

\[
= s^2 Var(D),
\]

which is the variance of R from Eq. 4.15.

4.7 The proof that "R(t) is a beta(a+1, b) distribution given that R is a beta(a, b) distribution for perfectly reliable forecasts."

The density function of the conditional distribution R(t) given that R is a beta(a, b) distribution can be determined by using Bayes' rule:

\[
f_{R(t)|R(t)}(R_t) = \frac{f_{R(t)}(1) f_{R(t)}(R_t)}{f_{R(t)}(1)}. \tag{A4}
\]

Since the density functions of D and D(t) are both discrete, we have

\[
f_{R(t)}(1) = Pr\{D=1\} = E(D) \]

\[
f_{D(t)=R(t)}(1) = Pr\{D=1|R=t\}.
\]

Under the assumption of perfect reliability (Eq. 4.12), we have the following relationship:

\[
f_{D(t)=R(t)}(1) = Pr\{D=1|R=t\} = R_t.
\]

Eq. A4 can now be written as
\[ f_{R|D=1}(R_t) = \frac{R_t}{E(R)} f_R(R_t) \]

Assuming \( R \) to be a beta(a,b) distribution, we obtain

\[ f_{R|D=1}(R_t) = \frac{R_t}{E(0) E(a,b)} R_t^{a-1} (1-R_t)^{b-1}, \quad (A5) \]

which is proportional to \( R_t^a (1-R_t)^{b-1} \). This result in Eq. A5 is sufficient to indicate that the conditional distribution of \( R|D=1 \) is a beta(a+1,b) distribution.

4.8 The proof that "\( R|D=0 \) is a beta(a,b+1) distribution given that \( R \) is a beta(a,b) distribution for perfectly reliable forecasts."

From Bayes' rule,

\[ f_{R|D=0}(R_t) = \frac{f_{D|R=R_t}(R_t) f_R(R_t)}{f_D(0)} , \]

\[ = \frac{(1 - \Pr \{ D=1 \mid R=R_t \}) f_R(R_t)}{1 - \Pr \{ D=0 \} \} \]

\[ = \frac{1 - R_t}{1 - E(0) E(a,b)} R_t^{a-1} (1-R_t)^{b-1} \]

\[ = \frac{1}{(1-E(0)) B(a,b)} R_t^{a-1} (1-R_t)^b . \]

Thus, the conditional distribution of \( R|D=0 \) is a beta(a,b+1) distribution.
4.9 Beta distribution of $R$ and $R-D$ for various values of the skill score $s^*$. $\circ$ denotes the mean of each branch of the beta distribution. $\ast$ denotes the point of $E(D)$ or $E(0)-1$.

(a) $s^* = 0.1$, $E(0) = 0.3$. 

**Diagram:**

- Beta distribution of $R$ for $s^*=0.1$ with $E(0)=0.3$.
  - Beta($2.70, 6.30$) distribution.
  - Beta($3.70, 6.30$) distribution.
  - Beta($2.70, 7.30$) distribution.

- $RIO=1$ and $RIO=0$.
(b) $s^* = 0.2$, $E(0) = 0.3$.  

\[ \text{BETA}(1.20, 2.80) \text{ DIST.} \]

\[ \text{BETA}(2.20, 2.80) \quad \text{BETA}(1.20, 3.80) \]
(c) $s^* = 0.3$, $E(0) = 0.3$.

$\beta(0.70, 1.63)$ DIST.

$\beta(1.70, 1.63)$  $\beta(0.70, 2.63)$
(d) \( s^* = 0.4, \ E(D) = 0.3 \).

\[
\text{BETA}(0.45, 1.05) \text{ DIST.}
\]

\[
\text{BETA}(1.45, 1.05) \quad \text{BETA}(0.45, 2.05)
\]

R.ID = 1 \quad \text{R.ID = 0}
(e) $s^* = 0.5$, $E(0) = 0.3$.

BETA(0.30, 0.70) DIST.

BETA(1.30, 0.70)  BETA(0.30, 1.70)

R10 = 1  R10 = 0
(f) $s^* = 0.6, \ E(D) = 0.3$.

$\beta(0.20, 0.47)$ DIST.

$\beta(1.20, 0.47)$  $\beta(0.20, 1.47)$
(g) $s^* = 0.7$, $E(0) = 0.3$. 

BETA(.13, .30) DIST.

BETA(1.13, .30)  BETA(.13, 1.30)

R10D=1  R10D=0
(h) $s^* = 0.8$, $E(0) = 0.3$.

BETA(.08, .18) DIST.

BETA(1.08, .18)  BETA(.08, 1.18)

R10 = 1  R10 = 0
(i) \( s^* = 0.9, \ E(0) = 0.3 \).

BETA(0.03, 0.08) DIST.

BETA(1.03, 0.08)    BETA(0.03, 1.08)
5.1. The proof that \( \bar{\bar{D}}(\bar{u}-\bar{\bar{D}})^* = 0.5 - (0.5 \bar{u} - \bar{\bar{D}}) \cdot (0.5 \bar{u} - \bar{\bar{D}})^* \).

\[
0.5 - (0.5 \bar{u} - \bar{\bar{D}}) \cdot (0.5 \bar{u} - \bar{\bar{D}})^* = 0.5 - 0.5 \cdot 0.5 \cdot \bar{u}^* + 2 \cdot \bar{u}^* \cdot 0.5 \cdot \bar{\bar{D}}^* - 0.5 \cdot \bar{\bar{D}}^*,
\]

\[
= 0.5 - (0.5, 0.5) x (0.5) + 0.5 x (1) \cdot \bar{\bar{D}}^* - \bar{\bar{D}}^*,
\]

\[
= \bar{\bar{D}}^* - \bar{\bar{D}}^*,
\]

\[
= \bar{\bar{D}}(\bar{u}-\bar{\bar{D}})^*.
\]

5.2. The proof that "the distance between \( \bar{\bar{D}}_t \) and the no-skill line is proportional to the skill \( s_t \)."

The skill is proportional to:

\[
RS_t - RE_t = (RS_t - d^2) + (d^2 - RE_t),
\]

\[
= (b+a)^2 - (c-a)^2.
\]

By using the identity

\[
x^2 - y^2 = (x+y)(x-y),
\]

\[
RS_t - RE_t = (b+a-c-a)(b+a-c+a),
\]

\[
= (b+c)(b-c+2a),
\]

\[
= 2(b+c)a. \quad (\text{since } b=c)
\]

Thus, \( RS_t - RE_t \) is proportional to "a", which is the distance between the point \( \bar{\bar{D}}_t \) and the no-skill line in the figure. Since skill \( s_t \) is proportional to \( RS_t - RE_t \), it is also proportional to "a".