Sensitivity of Verification Scores to the Classification of the Predictand

HARALD DAAN

Royal Netherlands Meteorological Institute, De Bilt, The Netherlands

(Manuscript received 18 February 1984, in final form 2 January 1985)

ABSTRACT

In the practice of forecast verification, the results of applying scoring rules appear to depend on the way the predictand is classified. This paper contains an examination of the sensitivity of six scoring rules to the classification. The approach is purely theoretical, in a sense that a Gaussian model for both forecasts and observations is designed. Scoring results for this model are calculated for different scoring rules and different classifications.

The results appear to favor the Ranked Probability Score (RPS), which is almost insensitive to the classification. Further, categorical scoring rules show a better performance in this respect than probabilistic scoring rules, except for the RPS. The use of the other three scoring rules (for probability forecasts) should not be recommended for the verification of forecasts of ordered predictands; that is, in case the classification involves more than two classes.

1. Introduction

Results of forecast verification by means of scoring rules should provide a measure of some attribute of the forecasts. The attribute of concern may be, for instance, accuracy or skill. In practice scoring results appear to depend on other conditions too. Particularly, the frequency distribution of the predictand concerned, and the way it was classified, play a role. Therefore, verification figures of forecasts for different predictands generally are not comparable, even in the case that the same scoring rule was applied.

In practical experiments, the use of different scoring rules for the same predictand also appears to provide a wide variety of results. Daan and Murphy (1982) presented verification figures of 2570 experimental probability forecasts of wind speed. For this sample the results were:

Ranked Probability Skill Score:	13.9%
Probability Skill Score:	2.1%
Logarithmic Skill Score:	4.6%.

Some skill scores also seem to depend rather strongly on the classification of the predictand. In the case mentioned above, the predictand was divided by three thresholds into four classes. A coarser classification in two classes, by maintaining only the middle threshold, would have resulted in a Brier skill score of 10.8%, quite different from the 2.1% result.

Not only the number of classes may play a role, but also the nature of the classification; that is, the way the thresholds are divided over the scale of the predictand. The division may be balanced with respect to the interval width (the distance between subsequent thresholds), or with respect to the frequency of each class, or even be quite unbalanced.

In summarizing, we find that verification results may depend on:

- the skill of the forecasts,
- the frequency distribution of the predictand,
- the scoring rule that was used,
- the number of classes, and
- the nature of the classification.

It is evident that skill scores preferably should reflect the skill of forecasts only. For that reason a study was dedicated to the sensitivity of six selected skill scores to the classification of the predictand. The approach taken to the problem was purely theoretical. A model has been designed, describing the predictand and its classification (Section 2) and the (probability) forecasts (Section 3). Section 4 contains a summary of the skill scores that were selected for examination. For each skill score and for two types of classification, diagrams were plotted, recording the scoring result as a function of a quality measure of the forecasts and of the number of classes (Section 5). From these graphs, finally, conclusions are drawn concerning the representativeness of the scoring results (Section 6).

2. Modeling of the predictand and its classification

a. The predictand

The predictand is assumed to be a one-dimensional quantity. That is, each observation can be characterized by a real number. Further, the climatological frequency distribution is assumed to be a Gaussian function, with a mean value of 0, and a standard

deviation of 1. In this paper a Gaussian function generally will be denoted as

$$g(x|\mu, \sigma), \tag{2.1}$$

where x denotes the argument, μ the mean value, and σ the standard deviation. The integral of this function over the interval (x_1, x_2) will be denoted by

$$G(x_1, x_2|\mu, \sigma). \tag{2.2}$$

b. Number of classes

The number of classes is denoted by T. Six values of T will be used, given by

$$T_r = 2^r \ (r = 1, 2, 3, 4, 5, 6). \tag{2.3}$$

Consequently, the number of classes is defined by the classification parameter r. Each class (except for the lowest class) corresponds with an half-open interval, the lower bound being included, the upper bound excluded. Further, the classifications are defined in such a way, that each class of classification r includes exactly two contiguous classes of classification $r + 1$, as shown in Fig. 1.

The parameter t denotes the rank order of a class, taking values from 0 through $T - 1$.

The climatological frequency of class t will be denoted by c_t. The parameter C_t will represent the climatological frequency of occurrence of classes, smaller than or equal to t:

$$C_t = \sum_{j=0}^{t} c_j. \tag{2.4}$$

c. Nature of the classification

Two basically different systems of classification have been involved in the model.

1) EQUIDISTANT (CONSTANT WIDTH) CLASSIFICATION

Here all classes, except for the first and last class, are intervals of an equal width. This width depends on the number of classes in the following way:

$$\text{width} = 8/T_r$$

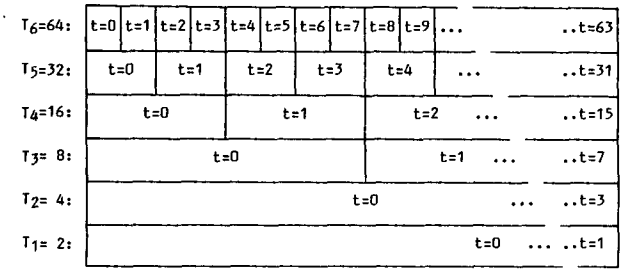

FIG. 1. Definition of the classifications for each of six values of r.

The parameters b_{rt} and b_{rt+1} will denote the lower and upper bound, respectively, of class t in classification r. The bounds of the classes are defined as follows:

$$b_{rt} = -4 + 8t/T_r \quad \text{for} \quad t = 1, \cdots, T - 1,$$

$$b_{r0} = -\infty \quad \text{and} \quad b_{rT} = +\infty. \tag{2.5}$$

Both of the outer classes have an infinite width (nevertheless, their frequency of occurrence is extremely small).

2) EQUIFREQUENT (CONSTANT FREQUENCY) CLASSIFICATION

Here the classes are chosen in such a way that the climatological frequencies c_t of all classes t are equal to $1/T$. In the vicinity of zero, the interval width of the classes is very small (for finer classifications about ⅓ of the width in the equidistant case), and it grows larger for greater departures from zero. The bounds are defined by

$$G(b_{rt}, b_{rt+1}|0, 1) = c_t = 1/T_r. \tag{2.6}$$

3. Modeling of the forecasts

a. General description

The forecasts are assumed to be derived directly from a judgmental probability distribution. The latter is represented by a Gaussian function with a mean value m' and a standard deviation s'. For the computations, it was necessary to limit the number of possible forecasts. Therefore, the number of possible judgments was restricted, by permitting only 32 discrete values of m'. The subscript i will be used to identify the judgment; m'_i is the mean of the probability distribution associated with judgment i (i taking values from 1 through 32). The standard deviation s' was assumed to be invariant with respect to i. The probability of occurrence of class t, based on the ith judgment, is denoted by

$$p_{it} = G(b_{rt}, b_{rt+1}|m'_i, s'). \tag{3.1}$$

The cumulative probability of occurrence of a class smaller than or equal to t is denoted by

$$P_{it} = G(-\infty, b_{rt+1}|m'_i, s'). \tag{3.2}$$

The forecasts may be biased; that is, the frequency distribution f_i of observations after a forecast based on judgment i is not necessarily identical to the forecast probability distribution p_i. We will assume that f_i (again) is Gaussian, with a mean value m_i, and a standard deviation s. Then, the observed frequency of class t, after forecasts based on judgment i, is denoted by

$$f_{it} = G(b_{rt}, b_{rt+1}|m_i, s), \tag{3.3}$$

and the cumulative frequency of occurrence of a class smaller than or equal to t is

$$F_{it} = G(-\infty, b_{t+1}|m_i, s). \quad (3.4)$$

b. Definition of judgments

The frequency of issue of each of the 32 possible forecasts is not arbitrary. For each possible forecast, a frequency distribution of observed values is defined, and we should require that the total of these 32 distributions equals the climatological frequency distribution. This was achieved by means of the following procedure. For the 32 possible values of m, the lower equifrequent class bounds b_{6t} for odd values of t were chosen. Consequently m_i is defined by

$$m_i = b_{6,2i-1} \quad (i = 1, 2, 3, \cdots, 32). \quad (3.5)$$

Then a weighting function $W(i)$ was defined, representing the frequency of occurrence of judgment i:

$$W(i) = G[b_{6,2i-2}, b_{6,2i}|0, (1 - s^2)^{1/2}], \quad (3.6)$$

where (again) the bounds b are derived from the $r = 6$ ($T = 64$) equifrequent classification. With these definitions, the requirement that:

$$\sum_{i=1}^{32} W(i) \cdot f_{it} = c_t \quad (3.7)$$

is approximately met.

c. Quality of the forecasts

If the forecasts are unbiased, that is, if $p_{it} = f_{it}$ for all i and t, then evidently s is a measure of the skill of the forecasts. In particular, if $s = 1$, then the forecast probabilities are identical to the climatological frequencies, and the forecasts are completely unskilled. On the other hand, if $s = 0$, then the forecasts are perfect. In the case the forecasts are biased the relationship is weaker, but still existing. We will introduce an attribute "quality" of the forecasts here, to be represented by a parameter q, with

$$q = 10(1 - s). \quad (3.8)$$

In the model we will use q in order to generate forecasts with different skill. Scoring results will be calculated for integer values of q, from 0 through 10. It should be noted that q is not assumed to be a quantitative measure of skill. On the other hand, when different samples of forecasts are equally biased, then skill is uniquely defined by q.

Nota bene. As opposed to s, the quantity q has a positive orientation; that is, higher values of q correspond with higher quality.

d. Bias in the forecasts

Bias in the forecasts may appear in two forms: 1) overconfidence and underconfidence, represented by differences between s' and s, and 2) systematic errors, represented by differences between m' and m. In this study only two cases are taken into account.

- Unbiased forecasts:

$$m' = m \quad \text{and} \quad s' = s. \quad (3.9)$$

- Biased forecasts: here a (quite arbitrary) bias is introduced, defined by:

$$m' = m - 0.3s \quad \text{and} \quad s' = 0.8s. \quad (3.10)$$

This bias represents a 25% overconfidence on the part of the forecaster, and a systematic error of 0.3 times the standard deviation.

4. The skill scores under examination

The sensitivity to the classification was originally observed in specific scoring rules for probability forecasts. Hence all the known scoring rules for this type of forecasts were involved in the study: the Probability score, the Ranked Probability Score, the Logarithmic Score and the Spherical Scoring Rule.

In effect, the study could easily be extended to scoring rules for categorical forecasts. The latter can be divided in two types: point-estimate statements, if only one value (or class) is forecast, and alternative statements, when each class is either forecast or not forecast. The first type of forecasts is verified by some measure of distance between forecast and observed class. Verification of the second type is based on the frequency of hits. Each of these verification schemes is represented in the study: the Mean Square Error Skill Score for point-estimate forecasts, and the Performance Index for alternative forecasts.

In the following, the six skill scores to be examined will be introduced. All scoring rules are formulated as skill scores; that is, the definition is such that a purely climatological forecast yields a zero result, whereas perfect forecasting is rewarded by a result 1. For each skill score, and for 6 values of r and 11 values of q, the expected scoring result (based on observed frequencies f_{it} for each judgment i) had to be calculated. These expected values will be denoted by the symbol E. In the introduction of the skill scores, the formulae for calculating these expected scoring results will be recorded as well. A separate Subsection 4g is devoted to the derivation of these formulae.

a. The mean square error skill score (MSE)

The MSE is a verification measure for point-estimate forecast statements. It is based on the squared distance between forecast value and observed value. In the verification scheme there is basically no need for a classification of the predictand. In this study, however, we will adopt the version that was suggested by Vernon (1953). This implies that the original scale

of the values of the predictand is replaced by a new discrete scale, involving the rank order of the classes. In this context MSE may be defined as follows:

$$\text{MSE} = 1 - (t_p - t_o)^2/E(t_c - t_o)^2, \quad (4.1)$$

where

t_p is the forecast class (the class to which the forecast mean value m' belongs),
t_o is the observed class,
t_c is the class to which the climatological mean (that is: 0) belongs: $t_c = T/2$, and
E denotes the expected value of $(t_c - t_o)^2$, which is a climatological constant.

The expected value of MSE may be evaluated from

$$E(\text{MSE}) = 1 - \sum_{i=1}^{32} \{W(i) \cdot \sum_{t=0}^{T-1} [f_{it} \cdot (t_{ip} - t)^2]\} / \sum_{t=0}^{T-1} [c_t \cdot (T/2 - t)^2]. \quad (4.2)$$

b. The performance index (PERF)

The Performance index (PERF; Hanssen and Kuipers, 1965) may be considered as a good representative of scoring rules for alternative forecasts (yes/no forecasts). The directive to the forecaster for this scoring rule is that a class should be forecast when its judgmental probability exceeds the climatological frequency; otherwise the class should not be forecast. The score reads:

$$\text{PERF} = \frac{\sum_{t=0}^{T-1} [\beta_t (o_t - c_t)]}{1 - \sum_{t=0}^{T-1} (c_t^2)}, \quad (4.3)$$

where

$\beta_t = 1$ if class t was forecast (if $p_t > c_t$), otherwise $\beta_t = 0$;
$o_t = 1$ if class t was observed, otherwise $o_t = 0$.

The expected value of PERF may be calculated from

$$E(\text{PERF}) = \sum_{i=1}^{32} \{W(i) \cdot \sum_{t=0}^{T-1} [\beta_{it} (f_{it} - c_t)]\} / [1 - \sum_{t=0}^{T-1} (c_t^2)], \quad (4.4)$$

where $\beta_{it} = 1$ if $G(b_{rt}, b_{rt+1} | m_i', s') > c_t$, otherwise $\beta_{it} = 0$.

c. The probability skill score (PROB)

The Probability Skill Score (PROB) considered here is a transformation of the Probability Score (Brier, 1950) and can be expressed as follows:

$$\text{PROB} = 1 - \sum_{t=0}^{T-1} (p_t - o_t)^2 / [1 - \sum_{t=0}^{T-1} (c_t^2)]. \quad (4.5)$$

The expected value of PROB is:

$$E(\text{PROB}) = 1 - \sum_{i=1}^{32} [W(i) \cdot \sum_{t=0}^{T-1} (p_{it}^2 - 2 p_{it} f_{it} + f_{it})] / [1 - \sum_{t=0}^{T-1} (c_t^2)]. \quad (4.6)$$

d. The information index (INFO)

This skill score is a slight modification of the Information Ratio, developed by Holloway and Woodbury (1955). The formula, used here, reads:

$$\text{INFO} = 1 - \ln(p_o) / \sum_{t=0}^{T-1} [c_t \cdot \ln(c_t)], \quad (4.7)$$

where the subscript o refers to the observed class. The expected value is:

$$E(\text{INFO}) = 1 - \sum_{i=1}^{32} \{W(i) \cdot \sum_{t=0}^{T-1} [f_{it} \cdot \ln(p_{it})]\} / \sum_{t=0}^{T-1} [c_t \cdot \ln(c_t)]. \quad (4.8)$$

e. The ranked probability skill score (RPS)

This index is a linear transformation of the RPS, as it was designed originally by Epstein (1969) and reformulated by Murphy (1971):

$$\text{RPS} = 1 - \sum_{t=0}^{T-1} (P_t - O_t)^2 / \sum_{t=0}^{T-1} [C_t \cdot (1 - C_t)], \quad (4.9)$$

where P_t and O_t denote the probability and observation (yes = 1, no = 0), respectively, of classes smaller than or equal to t. In the same way F_{it} is defined as the frequency of observation of classes smaller than or equal to t, associated with forecast i. Now the expected value of RPS may be calculated from:

$$E(\text{RPS}) = 1 - \sum_{i=1}^{32} [W(i) \cdot \sum_{t=0}^{T-1} (P_{it}^2 - 2 P_{it} F_{it} + F_{it})] / \sum_{t=0}^{T-1} [C_t \cdot (1 - C_t)]. \quad (4.10)$$

f. The spherical scoring rule (SPHER)

This score was mentioned in meteorological literature by Winkler and Murphy (1968). The formula is transformed into a skill score:

$$\text{SPHER} = \frac{p_o / [\sum_{t=0}^{T-1} (p_t^2)]^{1/2} - c_o / [\sum_{t=0}^{T-1} (c_t^2)]^{1/2}}{1 - [\sum_{t=0}^{T-1} (c_t^2)]^{1/2}}. \quad (4.11)$$

The subscript o, again, refers to the observed class. The expected value of SPHER is

$$E(\text{SPHER}) = \sum_{i=1}^{32} \langle W(i) \cdot \{ \sum_{t=0}^{T-1} (p_{it} f_{it}) / [\sum_{t=0}^{T-1} (p_{it}^2)]^{1/2} - [\sum_{t=0}^{T-1} (c_t^2)]^{1/2} \} \rangle / \{1 - [\sum_{t=0}^{T-1} (c_t^2)]^{1/2}\}. \quad (4.12)$$

g. Derivation of the formulae for expected scoring results

A discussion of the definitions of the skill scores is beyond the scope of this paper. The reader is referred to the authors mentioned in the above subsections. For a comprehensive treatment of these skill scores, see Daan (1984).

The formulae for expected results are derived in the following way. A scoring rule V is a function of forecasts and observations:

$$V = V(\text{forecasts, observations}). \quad (4.13)$$

In the model used here the number of possible forecasts is 32 and the number of possible observations is T. The relative frequency of forecast i is given by $W(i)$, whereas the relative frequency of observed class o is given by f_{io}. Therefore, the expected value of V may be written in general as:

$$E(V) = \sum_{i=1}^{32} W(i) \cdot \sum_{o=0}^{T-1} f_{io} \cdot V(\text{forecast } i, \text{ observation } o). \quad (4.14)$$

We will give an example (for the Probability Skill Score) of the way this scheme is carried out. In the definition of PROB (4.5), the left-hand term (1) and the denominator of the right-hand term are constants. We will deal with the numerator of the right hand term only:

$$V = \sum_{t=0}^{T-1} (p_t - o_t)^2 = \sum_{t=0}^{T-1} p_t^2 - 2 \sum_{t=0}^{T-1} p_t \cdot o_t + \sum_{t=0}^{T-1} o_t^2. \quad (4.15)$$

As $o_t^2 = o_t$, and $o_t = 1$ for only one class (otherwise $o_t = 0$), we may write:

$$V = \sum_{t=0}^{T-1} p_t^2 - 2p_o + 1. \quad (4.16)$$

According to (4.14), the expected value of V is

$$E(V) = \sum_{i=1}^{32} W(i) \cdot \sum_{o=0}^{T-1} f_{io} \cdot (\sum_{t=0}^{T-1} p_{it}^2 - 2p_{io} + 1) = \sum_{i=1}^{32} W(i) \cdot [\sum_{t=0}^{T-1} p_t^2 - 2 \sum_{o=0}^{T-1} p_{io} f_{io} + \sum_{o=0}^{T-1} f_{io}]. \quad (4.17)$$

Now o may be replaced by t, and we arrive at (4.6).

5. Results and discussion

The results are recorded graphically in Figs. 2–5. In the diagrams, the parameter q (in the diagrams recorded as **Q**) is plotted on the abscissa, representing the "quality" of the forecasts. On the ordinate, the parameter r (in the diagrams recorded as **R**) indicates the number of classes. Figures 2 and 4 reflect an equidistant classification, whereas Figs. 3 and 5 refer to an equifrequent classification. In Figs. 2 and 3 the forecasts are unbiased, whereas in Figs. 4 and 5 a bias is introduced as described in Subsection 3d. The scoring results are recorded by isopleths in intervals of 10%.

In Table 1, a quantitative survey of the results of unbiased forecasts for an equifrequently classified predictand is recorded. These data correspond to the diagrams in Fig. 3.

The scoring results of perfect forecasts ($q = 10$, unbiased forecasts) are not always exactly equal to 1. This may be caused by the fact that the formula for the weighting function $W(i)$ is only a discrete approximation. Moreover, the limitation on the number of possible judgments is not in accordance with the concept of perfect forecasting.

In the introduction, the proposition was stated that a scoring rule should return scoring results and reflect only the attribute of concern of the forecasts (i.e., the skill). This implies that a scoring rule should not be sensitive to the classification of the predictand. Since a very coarse classification may not be expected to do complete justice to very skillful forecasts, this requirement may be restricted to finer classifications only. In any case, a dependence of the scoring result V on q only should be considered an advantage of the scoring rule of concern. In the graphs this can be tested by checking whether the isopleths are (approximately) vertical lines, at least in the upper part of the diagram.

From the graphs, we find that this requirement is met by far the best by the RPS. The other three scores for probabilistic forecasts (PROB, INFO, SPHER) are clearly very sensitive to the number of classes. Moreover, the latter tend to approach to zero for finer classifications, only partly depending on the forecast quality parameter q. Results for both of the categorical scoring rules (MSE and PERF) at least tend to verticals for finer classifications.

An explanation for the difference in behavior of the RPS versus other scoring rules for probability forecasts is not obvious. The fact that the latter scores are not sensitive to distance, as opposed to the RPS (Staël von Holstein, 1970) may be important. If so, then the desirability of this property is emphasized strongly by the results. On the other hand, the results for the Performance Index (which is not sensitive to distance either) seems not to agree well with this assumption.

In any case, probability scores seem to be sensitive to the magnitude of the probabilities involved. This

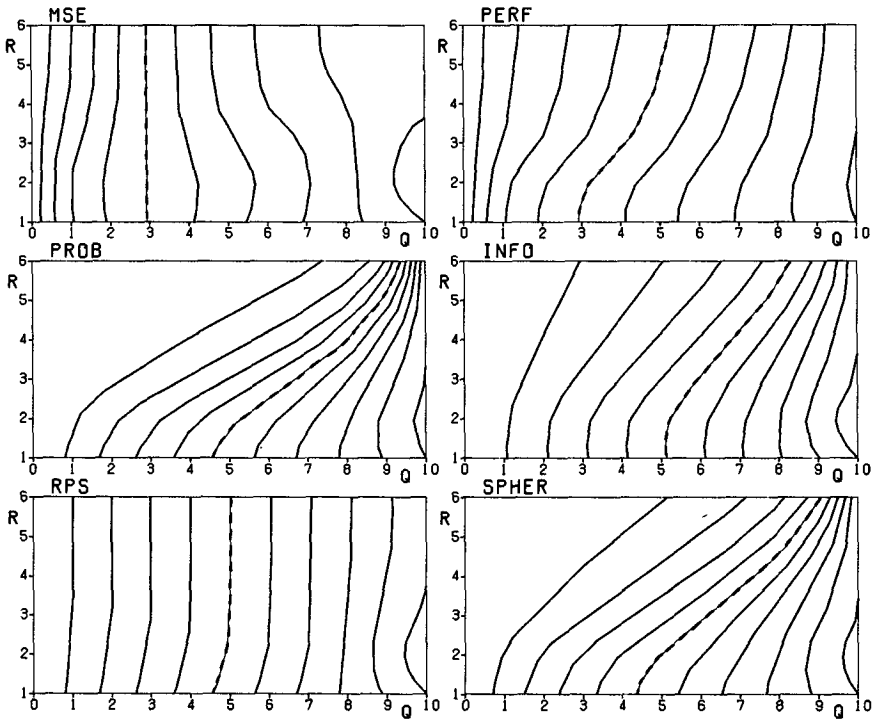

FIG. 2. Scoring results as a function of forecast quality and number of classes. The dashed line represents the 50% score isopleth. Abscissa: values of q(Q) from 0 to 10 (see Section 3c). Ordinate: values of r(R) from 1 to 6 (see Section 2b). The forecasts are unbiased (see Section 3d). The classification is equidistant (see section 2c).

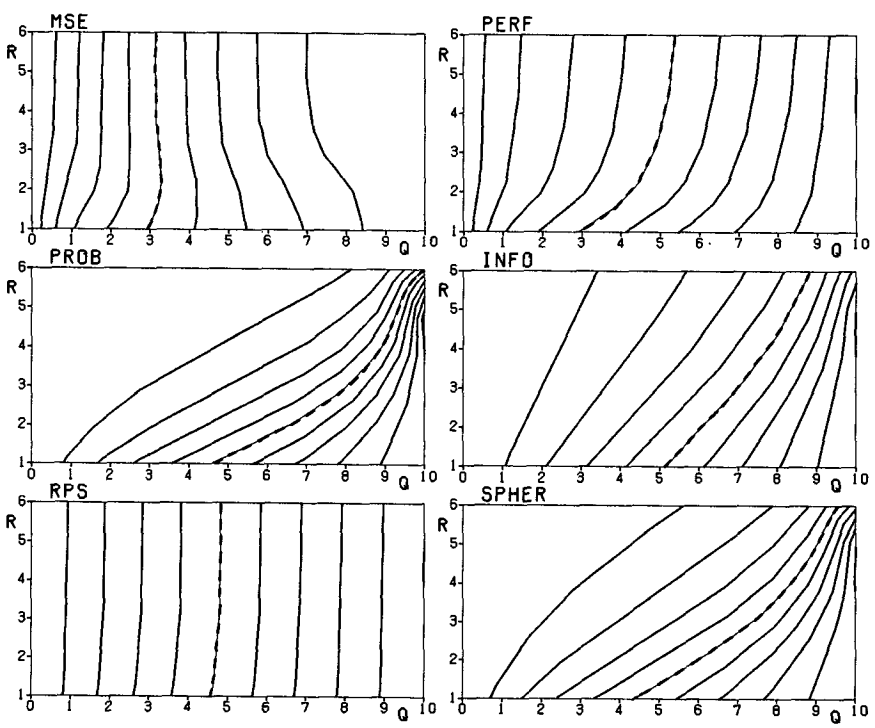

FIG. 3. As in Fig. 2 but the classification is equifrequent.

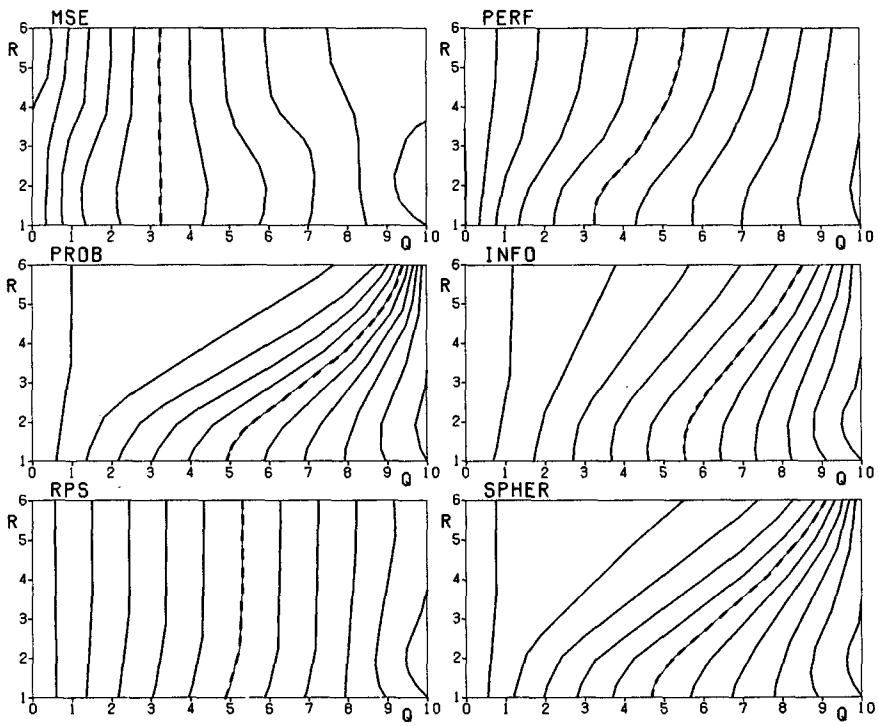

FIG. 4. As in Fig. 2 but the forecasts are biased.

magnitude decreases globally in inverse ratio with the number of classes, whereas for the RPS such a relationship does not exist.

Except for the RPS, scoring results for an equifrequent classification are generally lower than for an equidistant classification with the same number of

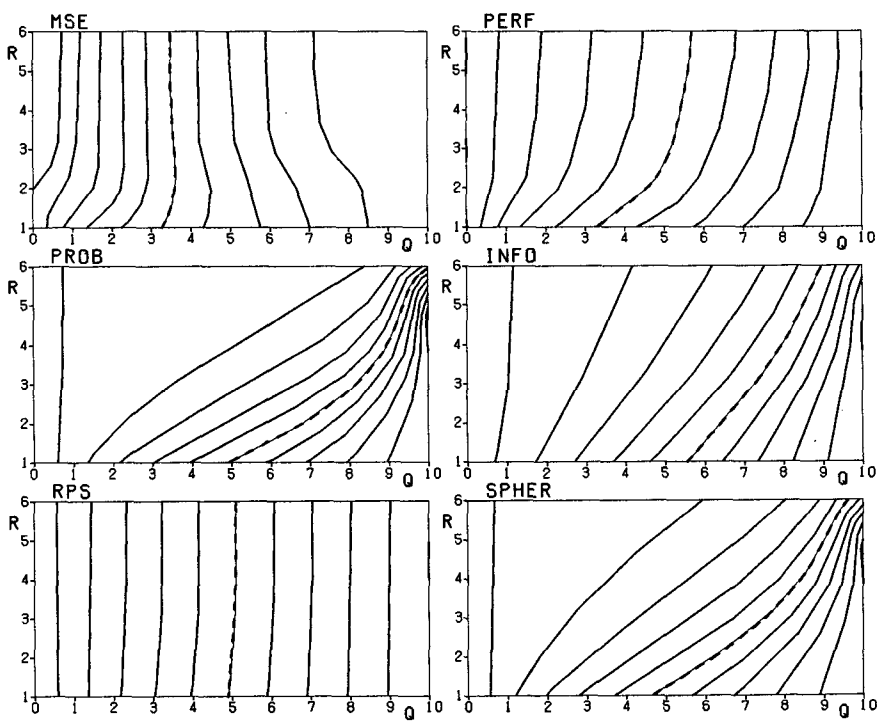

FIG. 5. As in Fig. 3 but the forecasts are biased.

TABLE 1. Scoring figures as a function of forecast quality and number of classes. The forecasts are unbiased and the classification is equifrequent.

	q										
r	0	1	2	3	4	5	6	7	8	9	10
						MSE					
6	0	17	33	48	61	73	83	90	95	99	100
5	0	17	33	48	61	73	83	90	96	99	100
4	1	17	33	48	61	73	82	90	95	98	100
3	2	18	33	47	60	71	81	88	93	97	100
2	3	22	34	47	58	68	76	83	89	95	100
1	3	29	41	51	59	67	74	81	87	94	100
						PERF					
6	2	16	24	32	39	47	55	64	75	86	98
5	2	16	24	32	40	48	56	65	75	87	100
4	2	17	25	33	41	49	57	67	77	88	100
3	2	18	27	35	43	51	60	69	79	90	100
2	2	20	30	39	48	56	65	74	82	91	100
1	3	29	41	51	59	67	74	81	87	94	100
						PROB					
6	0	0	1	1	2	3	4	6	9	19	49
5	0	1	2	3	4	6	8	12	18	35	100
4	0	1	3	5	8	11	15	21	31	53	100
3	0	3	6	10	15	20	27	36	49	72	100
2	0	6	12	19	26	34	44	55	69	84	100
1	0	12	23	34	44	54	63	73	82	91	100
						INFO					
6	0	3	5	9	12	17	22	28	38	53	83
5	0	3	6	10	15	20	26	34	45	63	100
4	0	4	8	13	18	24	31	41	53	72	100
3	0	5	10	16	23	30	39	50	64	81	100
2	0	6	13	21	29	39	49	60	73	87	100
1	0	9	19	29	38	49	59	69	79	90	100
						RPS					
6	0	11	21	32	42	52	62	71	81	90	98
5	0	11	21	32	42	52	62	71	81	91	100
4	0	11	22	32	42	52	62	71	81	91	100
3	0	11	22	32	42	52	62	71	81	91	100
2	0	11	22	33	43	53	62	72	81	91	100
1	0	12	23	34	44	54	63	73	82	91	100
						SPHER					
6	0	1	3	5	6	8	11	15	21	34	67
5	0	2	4	7	9	12	16	22	31	50	100
4	0	3	6	10	14	18	24	32	43	65	100
3	0	5	10	15	21	27	35	45	59	79	100
2	0	8	16	23	31	40	50	61	74	87	100
1	0	14	26	36	46	56	65	74	83	91	100

classes. The choice of an equifrequent classification instead of an equidistant classification seems essentially equivalent to an extension of the number of classes; that is, the results show much resemblance, apart from a shift along the vertical axis.

The slope of the isopleths seems somewhat more regular when the classification is equifrequent, than in the equidistant case. This holds particularly for low values of r (coarse classifications). For finer classifications the difference is negligible (that is, the difference in slope, not in scoring results).

The difference between unbiased forecasts and biased forecasts is small. Of course, scoring figures are lower in the latter case, but the slope of the isopleths seems independent of bias.

6. Conclusions

Six skill scores have been examined with respect to their sensitivity to the way the predictand was classified. For this purpose, a framework of forecasts and observations was constructed. The main results may be summarized as follows.

- The Ranked Probability Score is hardly sensitive to the classification of the predictand, and consequently, from this point of view, it should be considered the better scoring rule for probabilistic forecasts of an ordered predictand.
- The Mean Square Error and the Performance Index may be considered as useful scoring rules in this context, provided that the classification is fine enough. Because of their fundamental relationships, this (probably) holds for several other scoring rules too. The MSE represents a family of scoring rules, of which the Root Mean Square Error, the Mean Absolute Error, and the Variance also are members. The Performance Index is strongly associated with the Gringorten skill score (Gringorten, 1965) and with Heidke's skill score (Heidke, 1926).
- Other scores for probabilistic forecasts, as the Probability Score, the Information Ratio, and the Spherical Scoring Rule, appear to be highly sensitive to the definition of classes of the predictand. Without information on the nature of the classification, results of these scoring rules should not be judged to be representative measures of skill in forecasting. Therefore, the use of the latter scores in the verification of probability forecasts for an ordered predictand is not advisable. As the RPS is the average of a number of two-class Brier scores with different thresholds, this recommendation should be restricted to classifications into more than two classes. In other words, the Probability Skill Score, the Information Index, and the Spherical Scoring Rule should be applied to two-class predictands only.

As the results of this study are based on quite theoretical models of forecasts and observations, the question might arise whether conclusions can be transferred straightaway to operational forecast verification. For the negative results with respect to three

probabilistic forecast verification scores the answer should be that there are no firm grounds to suppose that these scores would be less sensitive to the classification in operational practice. On the other hand, it is possible that the sensitivity of the RPS to classification might increase in operational forecasting. This might be caused by forecast probability distributions departing strongly from a Gaussian function, and by very unbalanced classifications. For many weather elements the main operational thresholds are often located in extreme areas of the scale of the predictand; e.g. visibility and cloud base in aviation forecasts, or wind force in shipping forecasts. Nevertheless, the difference in sensitivity observed in the model between the RPS and the other three probability scores is impressive, and the recommendation to prefer the RPS in case of an ordered predictand is not likely to be affected in operational circumstances.

From the viewpoint of a user the application of the RPS instead of the Brier Probability Score is not illogical. The RPS is based on probabilities of exceeding certain thresholds, whereas in the Probability Score the probability of occurrence of values in certain intervals is involved. It is reasonable to assume that the practical use of weather forecasts in general corresponds better with the concept of thresholds than with that of intervals.

REFERENCES

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

Daan, H., 1984: Scoring rules in forecast verification. WMO, Geneva, PSMP Publ. Ser. No. 4, 62 pp.

——, and A. H. Murphy, 1982: Subjective probability forecasting in the Netherlands: Some operational and experimental results. *Meteor. Rundsch.*, **35,** 99–112.

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

Gringorten, I. I., 1965: A measure of skill in forecasting a continuous variable. *J. Appl. Meteor.*, **4,** 47–53.

Hanssen, A. W., and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Meded. Verh.*, **81,** 3–15.

Heidke, P., 1926: Berechnung des erfolges und der guete der windstaerkevorhersagen im sturmwarnungsdienst. *Geogr. Ann.*, **8,** 301–349.

Holloway, J. L., and M. A. Woodbury, 1955: Application of information theory and discriminant function analysis to weather forecasting and verification. Contract No. 551(07) Tech. Rep. 1, University of Pennsylvania, Institute for Cooperative Research, Philadelphia, PA 19174, 85 pp.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10,** 155–156.

Staël von Holstein, C-A. S., 1970: A family of strictly proper scoring rules which are sensitive to distance. *J. Appl. Meteor.*, **9,** 360–364.

Vernon, E. M., 1953: A new concept of skill score for rating quantitative forecasts. *Mon. Wea. Rev.*, **81,** 326–329.

Winkler, R. L., and A. H. Murphy, 1968: 'Good' probability assessors. *J. Appl. Meteor.*, **7,** 751–758.