Impacts of Feedback and Experience on the Quality of Subjective Probability Forecasts: Comparison of Results from the First and Second Years of the Zierikzee Experiment

ALLAN H. MURPHY

Department of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331

HARALD DAAN

Royal Netherlands Meteorological Institute, 3730 AE De Bilt, The Netherlands

(Manuscript received 19 January 1983, in final form 24 October 1983)

ABSTRACT

Subjective probability forecasts of wind speed, visibility and precipitation events for six-hour periods have been prepared on an experimental basis by forecasters at Zierikzee in The Netherlands since October 1980. Results from the first year of the experiment were encouraging, but they revealed a substantial amount of overforecasting (i.e., a strong tendency for forecast probabilities to exceed observed relative frequencies) for all events, periods and forecasters. Moreover, this overforecasting was reflected in a rapid deterioration in the skill of the forecasts as a function of lead time. In October 1981 the forecasters were given extensive feedback concerning their individual and collective performance during the first year of the experimental program. The purpose of this paper is to compare the results of the first and second years of the experiment.

Evaluation of the forecasts formulated in the first and second years of the Zierikzee experiment reveals marked improvements in reliability (i.e., reductions in overforecasting) from year 1 to year 2, both overall and for most stratifications of the results by event, period or forecaster. For example, the reliability of the forecasts increased for all events and periods and for three of the four forecasters. The improvements in reliability are reflected in substantial increases in the skill of the forecasts from year 1 to year 2, with overall skill scores for the second (first) year for the wind speed, visibility and precipitation forecasts of 25.4% (13.9%), 22.4% (12.4%) and 0.5% (−24.7%), respectively. These improvements in performance are attributed to the feedback provided to the forecasters at the beginning of the second year of the experiment and to the experience in probability forecasting gained by the forecasters during the first year of the program.

The paper concludes with a brief discussion of the results and their implications for probability forecasting in meteorology.

1. Introduction

Since October 1980, subjective probability forecasts of wind speed, visibility, and precipitation events have been prepared on an experimental basis by Royal Netherlands Meteorological Institute (KNMI) forecasters at Zierikzee. The forecasts specify the likelihood that certain critical threshold values of the respective elements will or will not be exceeded in six-hour periods. Results from the first year of this experiment were summarized in a recent paper by Daan and Murphy (1982) (hereafter referred to simply as DM). While these results were encouraging, they revealed a substantial amount of overforecasting (i.e., forecast probabilities greater than observed relative frequencies) for all elements and thresholds, all periods, and all forecasters. Moreover, although skill scores for the forecasts were positive for all elements in the first period, they decreased markedly to marginally positive or negative values by the second period. It should be noted that the forecasters at Zierikzee had no prior experience at probability forecasting and that they did not receive any formal feedback concerning their performance during the first year of the experiment.

In early October 1981 the forecasters at Zierikzee were provided with detailed feedback regarding their performance during the first 12 months of the experiment. Moreover, at this time it was decided to continue the experiment for a second year so that the effects of feedback and experience on forecaster performance could be investigated. The primary purpose of this paper is to compare the results of the first and second years of the experimental program. Some details concerning the experiment at Zierikzee are presented in Section 2, together with information regarding the feedback provided to the forecasters in October 1981. The methods used to assess the quality of the experimental results are described briefly in Section 3. The results of the first and second years of the experiment are evaluated and compared in Section 4. Section 5 consists of a discussion and conclusion.

2. The Zierikzee experiment

The primary objective of the Zierikzee experiment was to investigate whether forecasters were able to formulate reliable and skillful forecasts of the probability that certain critical threshold values of important me-

teorological elements such as wind speed, visibility, and precipitation would (or would not) be exceeded. Forecasts of wind speed and visibility are of particular significance at Zierikzee, since this station provides meteorological information in support of activities related to the construction of large dams along the southwest coast of The Netherlands (the so-called Delta Project). In the case of wind speed, critical values of 6.2, 10.8, and 16.0 m s^{-1} were selected, corresponding to probability forecasts of the wind exceeding Beaufort force 3, 5, and 7, respectively. For visibility, probabilities were assessed for the occurrence of values below 1, 4, and 10 km. Finally, the forecasters prepared probability forecasts for precipitation amounts exceeding two thresholds, 0.2 and 1.4 mm. For convenience, we denote wind speed by FF, visibility by VV, and precipitation by RR.

Experimental probability forecasts were prepared twice a day (at 0600 and 1600 GMT) for five consecutive six-hour periods (the first period began zero and two hours, respectively, after the forecast time). The forecasts were expressed in terms of a one digit number from 0 to 9, where the numbers 0, 1, . . . , 9 represent probabilities in the ranges 0.0–0.1, 0.1–0.2, . . . , 0.9–1.0, respectively. The data used to evaluate the experimental forecasts were taken from observations made at or near Zierikzee (for additional details, see DM).

The forecasters at Zierikzee who participated in the experiment all possessed considerable general experience in weather forecasting. However, these forecasters did not have any experience in expressing forecasts in terms of probabilities prior to October 1980. In fact, at that time, not even climatological probabilities of the relevant wind speed, visibility, and precipitation events were known to the forecasters. Moreover, while standard numerical guidance in the form of prognostic charts was available to the forecasters, they were not provided with objective probabilistic guidance forecasts. Finally, the forecasters did not receive any formal feedback concerning their performance during the first 12 months of the experimental period.

The authors visited the station at Zierikzee in early October 1981. The primary purpose of this visit was to discuss the results of the first year of the experiment with the forecasters. At this meeting, the forecasters were provided with individual and collective feedback concerning their performance. Specifically, for each element they received reliability diagrams for each forecaster, for each period, and for each threshold. (Note: See Sections 3 and 4 for descriptions and examples of reliability diagrams.) The nature and implications of these results were discussed with the forecasters in considerable detail. In addition, for each element the forecasters were given various quantitative scores that described the reliability, resolution, and skill of the experimental forecasts for each forecaster, for each period, and for all forecasters and periods combined. At the conclusion of the meeting, it was decided to continue the experiment for a second year so that the effects of both feedback (provided during the meeting) and experience (gained from the first year) on forecaster performance could be evaluated.

The results reported in this paper are based on probability forecasts formulated during the first year (5 October 1980–30 September 1981) and second year (1 October 1981–30 September 1982) of the experiment. A total of 514 (544) sets of forecasts were prepared during the first (second) year. Each set consists of forecasts for three elements for five periods or lead times, for a total of 7710 (8160) forecasts. Moreover, since each forecast contains probabilities corresponding to three threshold values for wind speed and visibility and to two threshold values for precipitation, a grand total of 20 560 (21 760) probabilities were assessed by the forecasters in the first (second) year of the experiment. The four forecasters affiliated with the station at Zierikzee made 501 (528) of the 514 (544) sets of forecasts during the respective years, with forecasters at De Bilt formulating the forecasts on the other 13 (16) occasions.

3. Methods of evaluation

The attributes of the experimental probability forecasts of primary interest here are accuracy, reliability, resolution, and skill. Accuracy represents the average degree of correspondence between individual forecasts and observations. The measure of accuracy employed in this study is the ranked probability score (RPS) (Epstein, 1969; Murphy, 1971). This scoring rule is a particularly appropriate measure of accuracy in situations involving forecasts that specify the probability that two or more different threshold values of an ordinal variable will or will not be exceeded. The RPS, unlike the Brier score (Brier, 1950) and some other familiar measures of the accuracy of probability forecasts, takes this underlying ordering of the values of the variable into account (see Murphy, 1970; Stael von Holstein, 1970). Note that the elements of concern in this paper are ordinal variables involving two (precipitation) or three (wind speed and visibility) threshold values.

In order to define the RPS in the simplest possible form, it is necessary to introduce some notation for the relevant forecasts and observations. Consider a sample of K forecasts and let N denote the number of threshold values used in defining the (overlapping) events of interest. Further, let R_{kn} denote the forecast probability of exceeding (or not exceeding) the nth threshold value of a particular element on the kth occasion. Similarly, let $D_{kn} = 1$ if the observed value of this element exceeds (does not exceed) the nth threshold on the kth occasion and $D_{kn} = 0$ otherwise. Using this notation, the average ranked probability score (RPS) can be written as

$$\overline{\text{RPS}} = \sum_{n=1}^{N} [(1/K) \sum_{k=1}^{K} (R_{kn} - D_{kn})^2]. \quad (1)$$

$\overline{\text{RPS}}$ in Eq. (1) is the mean square error of probability forecasts (as defined in this paper), and it has a range from zero (best possible score) to N (worst possible score).

Note that $\overline{\text{RPS}}$ in Eq. (1) also represents the sum of N scores, each of which is one-half of the Brier score for probability forecasts in a two-event situation corresponding to one of the N possible threshold values. Each of these Brier scores can be partitioned into three terms, under the assumption that the forecast probabilities (i.e., the R_{kn}) possess only a finite number, S say, of distinct values R^s ($s = 1, \ldots, S$) (see Murphy, 1973; Murphy and Daan, 1983). For the nth threshold value, this assumption allows the sample of K forecasts to be divided into S subsamples, each of which contains all the forecasts for which $R_{kn} = R^s$. Then,

$$\overline{\text{RPS}} = N(\text{UNC} + \text{REL} - \text{RES}), \quad (2)$$

where

$$\text{UNC} = (1/N) \sum_{n=1}^{N} \bar{D}_n(1 - \bar{D}_n),$$

$$\text{REL} = (1/N) \sum_{n=1}^{N} (1/K) \sum_{s=1}^{S} K^s(R^s - \bar{D}_n^s)^2,$$

$$\text{RES} = (1/N) \sum_{n=1}^{N} (1/K) \sum_{s=1}^{S} K^s(\bar{D}_n^s - \bar{D}_n)^2.$$

In these expressions, K^s is the number of forecasts in subsample s ($\sum_s K^s = K$), $\bar{D}_n^s = (1/K^s) \sum_k D_{kn}$ (the summation is taken over all values of k in subsample s), and $\bar{D}_n = (1/K) \sum_k D_{kn}$.

The UNC term in Eq. (2) is simply the variance of the observations in the sample. This term represents a measure of the uncertainty inherent in the observations, and its values range from zero (minimum uncertainty—the nth threshold is always exceeded and the n + 1st threshold is never exceeded for some n) to one-fourth (maximum uncertainty—the first threshold is not exceeded on exactly one-half of the K occasions and the Nth threshold is exceeded on the remaining occasions). It should be noted that the variance term does not depend on the forecasts in any way.

The REL term in Eq. (2) is the reliability term, which is the weighted mean squared difference between the forecasts and the corresponding observed relative frequencies over the S subsamples. This term, which is non-negative and vanishes only when the forecast probabilities and observed relative frequencies are equal for all subsamples, provides a quantitative measure of the reliability of the forecasts. This attribute of the forecasts also can be examined qualitatively by means of a reliability diagram in which the relative frequency of the events is plotted against the forecast probability for specific probability values. The broken line connecting these sample points is then compared with the diagonal 45° line representing perfect reliability (equality of probability and relative frequency).

The RES term in Eq. (2) is the resolution term, which is the weighted mean squared difference between the observed relative frequencies in the subsamples and the overall (i.e., sample) observed relative frequencies. This term also is non-negative, and it vanishes only when all the subsample relative frequencies are equal to the sample relative frequencies. Since the sign of the resolution term is negative, the larger this term the greater the accuracy of the forecasts (the opposite is true for the reliability term).

Skill is defined generally as the accuracy of the forecasts of interest relative to the accuracy of forecasts produced by some reference procedure. The reference procedure used here is a (constant) forecast of the overall sample relative frequency (i.e., the sample climatological probability). In this case, it is natural to define a skill score SS as follows:

$$\text{SS} = [1 - (\overline{\text{RPS}}/\overline{\text{RPS}}_c)] \times 100\%. \quad (3)$$

Such a skill score is positive (negative) when the forecasts of interest are more (less) accurate than climatological forecasts, with SS = 100% representing the best possible skill score.

It is relatively easy to show that

$$\overline{\text{RPS}}_c = \sum_{n=1}^{N} \bar{D}_n(1 - \bar{D}_n) = N(\text{UNC}) \quad (4)$$

(Murphy and Daan, 1984). Therefore, using Eqs. (2) and (4), Eq. (3) can be rewritten as

$$\text{SS} = [(\text{RES} - \text{REL})/\text{UNC}] \times 100\%. \quad (5)$$

Thus, the skill of the forecasts is positive (negative) if the magnitude of the resolution term is greater (less) than the magnitude of the reliability term.

4. Some results: Second year versus first year

The average forecast probabilities, observed relative frequencies, and overall biases (ratios of average forecast probabilities to corresponding observed relative frequencies) for each element and threshold value in the first and second years of the experiment are presented in Table 1. These observed relative frequencies are the so-called sample climatological probabilities for the respective periods. As noted in DM, the forecast probabilities exceeded the observed relative frequencies (i.e., the biases exceeded unity) for all combinations of elements and thresholds in the first year. In the context of probability forecasting, a bias value of unity is optimal; that is, the sum of the probabilities assigned to an event should be equal to the number of times

TABLE 1. Average forecast probabilities, observed relative frequencies, and overall biases for experimental probability forecasts for second year (544 forecasts for each element/threshold combination). Corresponding figures for first year are given in parentheses (514 forecasts for each element/threshold combination).

Element/threshold	Average forecast probability	Observed relative frequency	Overall bias
Wind speed (FF)			
FF > 6.2 m s^{-1}	0.732 (0.769)	0.753 (0.752)	0.97 (1.02)
FF > 10.8 m s^{-1}	0.321 (0.398)	0.299 (0.309)	1.07 (1.29)
FF > 16.0 m s^{-1}	0.093 (0.101)	0.051 (0.036)	1.82 (2.81)
Visibility (VV)			
VV < 1 km	0.082 (0.098)	0.027 (0.032)	3.04 (3.06)
VV < 4 km	0.202 (0.269)	0.171 (0.180)	1.18 (1.49)
VV < 10 km	0.422 (0.496)	0.444 (0.433)	0.95 (1.15)
Precipitation (RR)			
RR > 0.2 mm	0.268 (0.399)	0.198 (0.246)	1.35 (1.62)
RR > 1.4 mm	0.164 (0.251)	0.094 (0.114)	1.74 (2.20)

this event is actually observed. Substantial improvements in the bias values from the first year to the second year can be seen in Table 1 for seven of the nine element/threshold combinations. Larger bias values are associated with more severe (or, equivalently, less frequent) events in both years of the experiment, but this tendency clearly is less pronounced in the second year for the wind speed and precipitation forecasts.

A reliability diagram for all elements, thresholds, periods, and forecasters combined is presented in Fig. 1a. This diagram provides an overall evaluation of the reliability of the subjective probability forecasts formulated during the first and second years of the Zierikzee experiment (the points in the diagram are plotted at the respective midpoints of the ranges of the probability values; see DM). A comparison of the two curves reveals that the reliability of the second year's forecasts is as good or better than that of the first year's forecasts for all probability values. The improvement in reliability is particularly noteworthy for probability values greater than 0.50, although a moderate degree of overforecasting still existed in the second year for these larger probabilities. The tendency for overforecasting to increase as the probability value increased also is less pronounced during the second year of the experiment. The overall frequency of use of the various probability values in the first and second years of the experiment is depicted in Fig. 1b. This diagram indicates that probabilities smaller (larger) than 0.40 were used more (less) frequently during the second year than during the first year. Differences in the frequency of use of probabilities between the two years of the experiment are greatest for the lowest and highest probability values.

Reliability diagrams for the experimental probability forecasts of wind speed, visibility, and precipitation are presented in Figs. 2, 3, and 4, respectively. These diagrams contain reliability curves for the first and second years of the experiment for each threshold (periods and forecasters are combined). The reliability curves for the various element/threshold combinations in Figs. 2–4 all exhibit the same general behavior. Specifically, the reliability curves for the second year usually follow the 45° line more closely than the corresponding curves for the first year, thereby indicating that the former involve less overforecasting than the latter. Moreover, the differences between these curves (i.e., the amount of improvement) generally is greater for larger probability values. The amount of overforecasting increases, in general, as the sample climatological probabilities of the events decrease in both years of the experiment, with the amount of improvement from year 1 to year 2 increasing as these probabilities decrease. In summary, overforecasting exists for most element/threshold combinations in the second year (especially for forecasts associated with larger probability values), but the amount of overforecasting is markedly less than that in the first year. Relative frequency of use distributions corresponding to these reliability curves exhibited the same general behavior as the overall frequency of use curves in Fig. 1b, and these distributions have been omitted to conserve space.

Differences in reliability between the year 1 and year 2 forecasts also were examined as a function of lead time (reliability diagrams omitted). Reliability generally is better for forecasts associated with shorter lead times (in both years of the experiment), and improvements in reliability between the first and second years usually are greater for longer lead times. The frequency of use of very high and very low probabilities tended to decrease as lead time increased in both year 1 and year 2.

Individual reliability diagrams for the four forecasters at Zierikzee are presented in Fig. 5 (elements, thresholds, and lead times combined). Each diagram contains separate reliability curves for the first and second years of the experiment. These diagrams indicate that Forecaster B's reliability curves (Fig. 5b) show the greatest improvement; in fact, this forecaster improved the reliability of his forecasts for all probability values. Forecasters A and C (Figs. 5a and 5c, respectively) also noticeably improved the reliability of their forecasts for larger probability values. Forecaster D's curves (Fig. 5d) reveal little if any improvement in reliability from year 1 to year 2, but the reliability of his wind speed and precipitation forecasts was better than that of the other three forecasters *in the first year of the experiment.* All four forecasters exhibit a tendency to overforecast in both years, with the amount of overforecasting generally increasing as the probability values increase (except for the largest probability values). Differences among forecasters in terms of the frequency of use of the various probability values (diagrams omitted) are quite small—in general, these distributions are similar

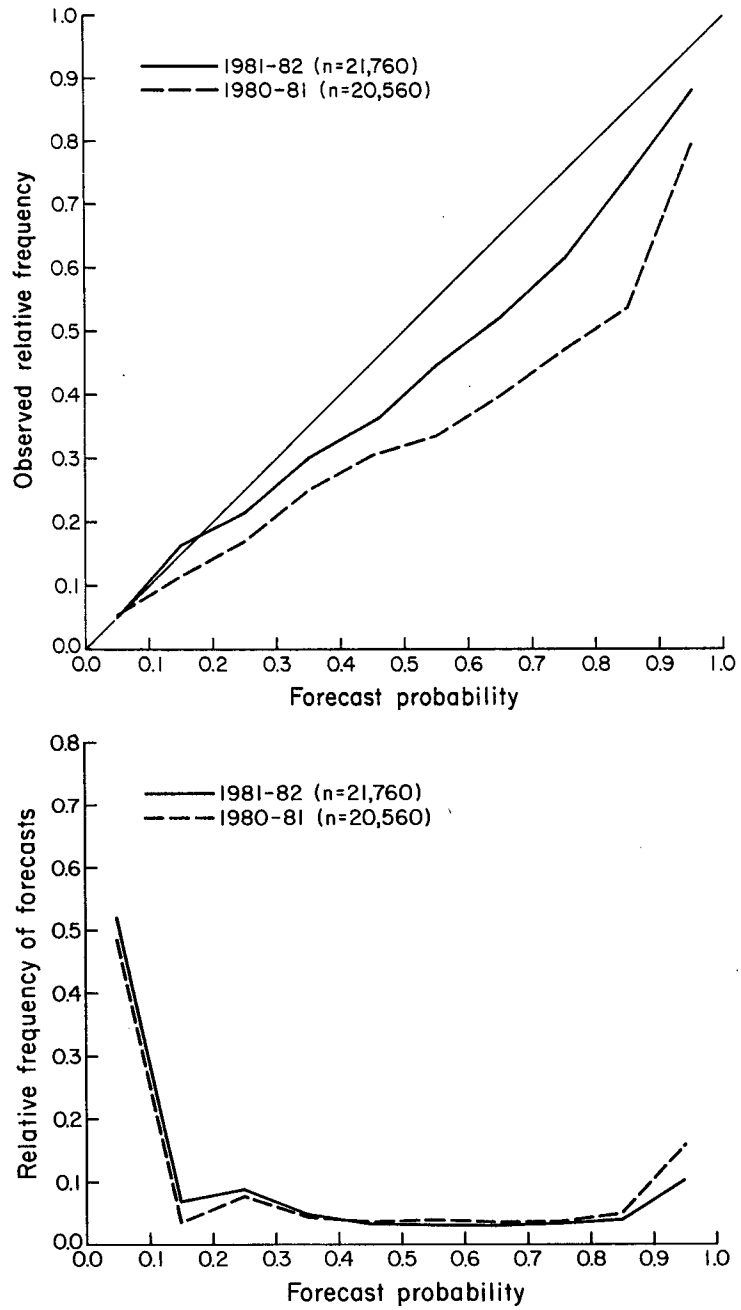

FIG. 1. (a) Overall reliability curves for subjective probability forecasts formulated in first and second years of Zierikzee experiment (all elements, thresholds, periods, and forecasters combined). (b) Overall relative frequency of use of probability values in first and second years of Zierikzee experiment.

in shape to the overall frequency of use distributions depicted in Fig. 1b.

Quantitative measures of performance for the experimental probability forecasts of wind speed, visibility, and precipitation are presented in Tables 2, 3, and 4, respectively, for the first and second years of the experiment (the scores for the first year appear in parentheses). The measures include the terms in the partition of the ranked probability score (RPS)—namely, the variance of observations (UNC), reliability (REL), and resolution (RES)—and the skill score (SS) based on the RPS (see Section 3 for definitions of these measures). In addition to the overall scores for each element, the results are stratified (separately) by threshold, period (or lead time), and forecaster. The overall results [part (a) of each table] indicate sub-

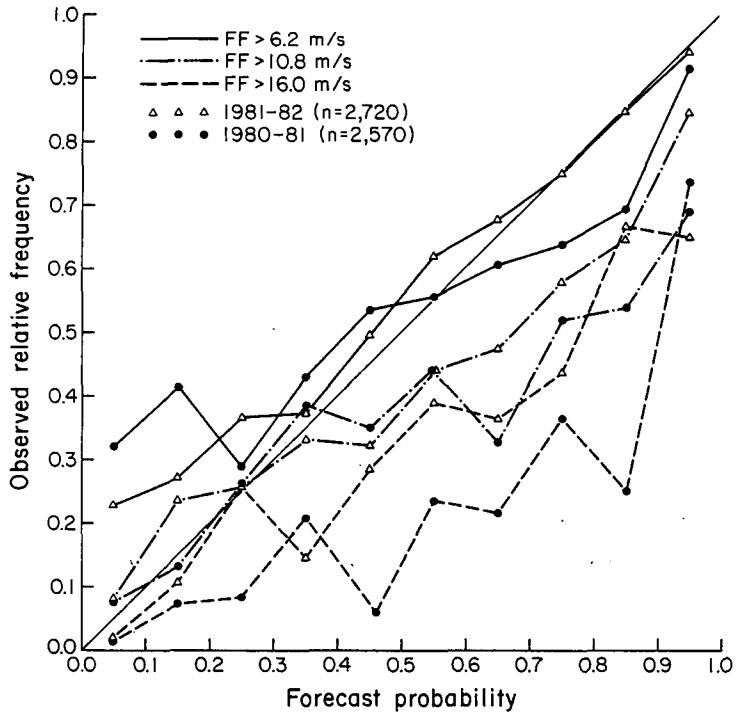

FIG. 2. Reliability curves for experimental probability forecasts of wind speed at three thresholds for years 1 and 2 (all periods and forecasters combined).

stantial improvements in skill from year 1 to year 2 for all three elements. Specifically, the skill of the wind speed and visibility forecasts in the second year is almost twice that in the first year, and the overall skill of the precipitation forecasts increases from a substantial negative value to a marginal positive value. It

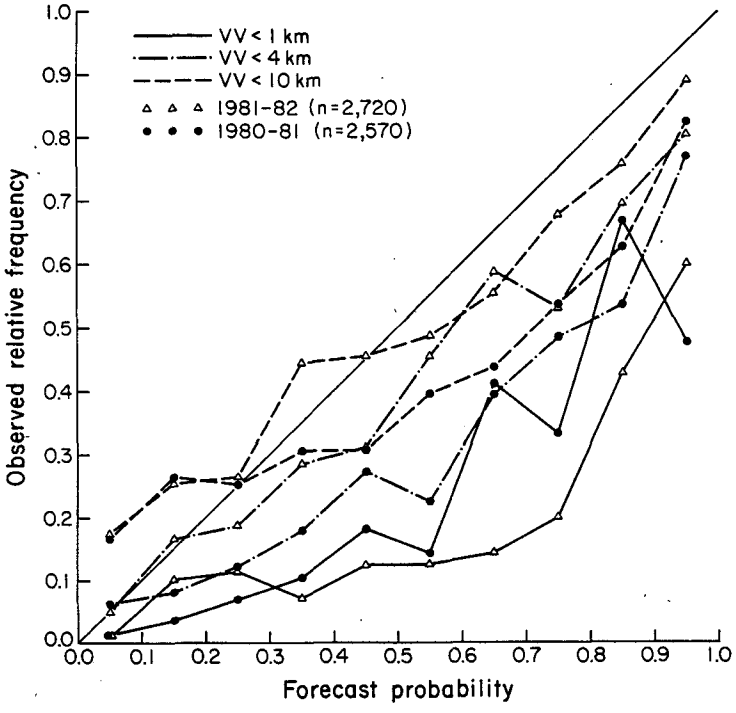

FIG. 3. Reliability curves for experimental probability forecasts of visibility at three thresholds for years 1 and 2 (all periods and forecasters combined).

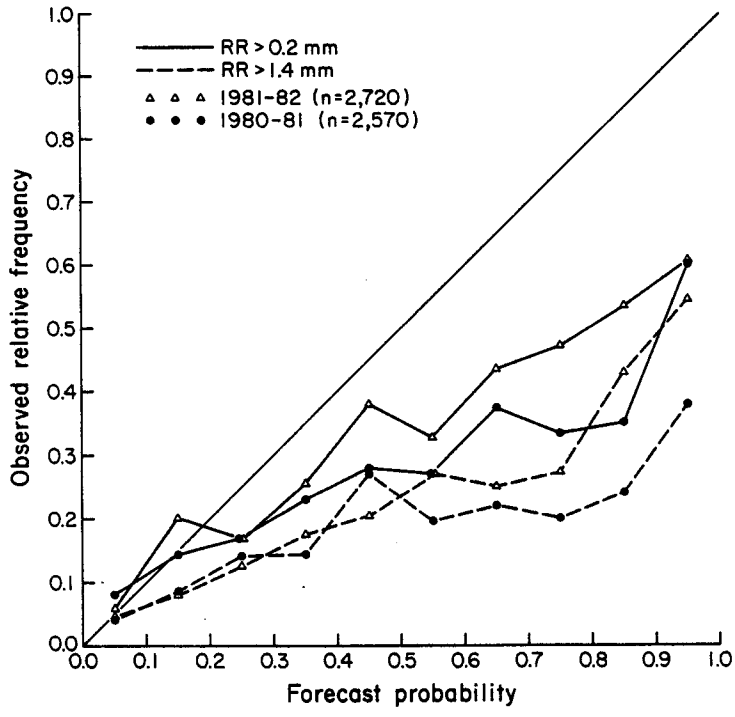

FIG. 4. Reliability curves for experimental probability forecasts of precipitation at two thresholds for years 1 and 2 (all periods and forecasters combined).

is of interest to note that these improvements are due largely to decreases in the magnitudes of the respective reliability terms from year 1 to year 2. Small improvements in resolution (i.e., increases in RES) also were recorded for the wind speed and visibility forecasts, but the magnitude of the resolution term actually decreased between year 1 and year 2 for the precipitation forecasts. In this regard, recall that SS = (RES − REL)/UNC [see Eq. (5)], so that the relative magnitudes of the reliability and resolution terms determine the sign of the skill score.

Tables 2–4 also indicate changes in UNC, REL, RES, and SS as a function of threshold [part (b)], lead time [part (c)], and forecaster [part (d)] for the respective elements. Increases in skill occurred from year 1 to year 2 for all element/threshold combinations except for VV < 1 km. Skill improved for all element/lead time combinations from 1980–81 to 1981–82, with a tendency for larger improvements to occur at longer lead times. Finally, three of the four forecasters showed moderate to substantial increases in skill between year 1 and year 2. These increases in skill were due largely to improvements in the reliability of the forecasts, since noticeable improvements in resolution occurred for only a few element/threshold, element/lead time, or element/forecaster combinations.

5. Discussion and conclusion

Overall, a comparison of the results for the first and second years of the Zierikzee experiment indicates substantial improvement in the reliability of the forecasts, whereas differences in resolution between year 1 and year 2 are relatively small. The improvements in reliability are reflected in marked increases in the skill of the experimental probability forecasts for all three elements. Moreover, stratification of the results for each element by threshold, period, and forecaster (separately) reveals improvements in reliability and skill from year 1 to year 2 for all element/threshold combinations, for all element/period combinations, and for all element/forecaster combinations except those involving Forecaster D. To illustrate these improvements in performance, skill scores (SS) and the skill score contributions of the REL and RES terms [see Eq. (5)] for each element are depicted in Fig. 6 as a function of lead time.

In order to place these results in the proper context, it is instructive to recall the explanations offered by DM for the overforecasting that occurred in the first year of the experiment. They attributed this overforecasting primarily to two factors: 1) the forecaster's lack of experience in probability forecasting (and, conversely, their extensive experience in categorical forecasting) and 2) a "value-induced" bias that frequently arises in situations such as the Zierikzee experiment when the forecasters recognize that the events of concern—and their forecasts—have a significant impact on users' activities and operations. In support of this explanation, it should be noted that substantial overforecasting has been observed in other probability forecasting experiments in which the forecasters (and/

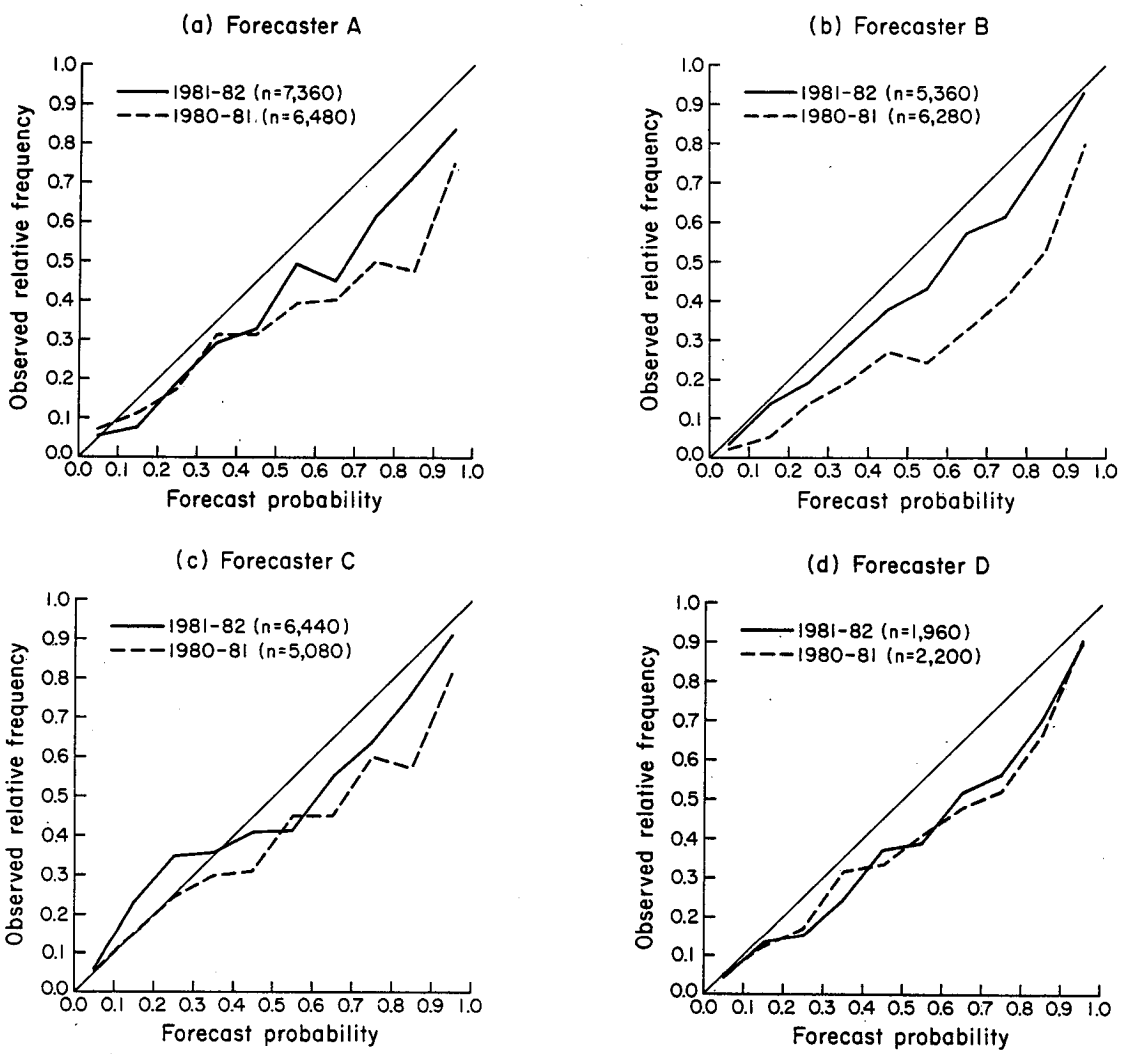

FIG. 5. Reliability curves for experimental probability forecasts for individual forecasters for years 1 and 2 (all elements, thresholds, and periods combined). (a) Forecaster A, (b) Forecaster B, (c) Forecaster C, (d) Forecaster D.

or the situations) possessed "characteristics" similar to those of the Zierikzee experiment (e.g., see Murphy *et al.*, 1982; Murphy and Winkler, 1982; Wallsten and Budescu, 1983).

We believe that the feedback given to the forecasters in October 1981—and the discussions that accompanied the provision of this feedback—were instrumental in enabling the forecasters to formulate more reliable and skillful forecasts in the second year of the experiment. In this regard, the feedback provided graphic and forecaster-specific results describing the nature and extent of the overforecasting that occurred in year 1. Moreover, the need to separate the process of formulating the forecasts from the process of using the forecasts was emphasized in the discussions with the forecasters, in an effort to reduce the effects of any value-induced bias. Of course, the general experience in probability forecasting gained by the forecasters during the first year of the experiment also undoubtedly had a beneficial effect (it is not possible to separate the effects of feedback and experience in this study). In any case, these results appear to represent a good illustration of the potential value of providing forecasters with feedback concerning their performance, especially in the early stages of a probability forecasting program. Of course, it generally would be desirable to provide forecasters with feedback at more frequent intervals than on an annual basis.

Consideration of the interpretations of the terms reliability and resolution provides some further insight into the implications of the results of this experiment. Reliability relates to a forecaster's ability to provide a proper "label" for the events of interest, where this label reflects the likelihood of occurrence of these events. Clearly, the forecasters at Zierikzee were more successful in year 2 than in year 1 at performing the

TABLE 2. Partition of average ranked probability score and skill score based on average ranked probability score for experimental probability forecasts of wind speed (FF) for second year: (a) overall; (b) by threshold; (c) by forecast period; and (d) by forecaster. Corresponding figures for first year are given in parentheses.

Stratification	Number of probabilities	Partition of ranked probability score: Variance of observations	Reliability	Resolution	Skill score (%)
(a) None	8160 (7770)	0.148 (0.145)	0.005 (0.016)	0.043 (0.036)	25.4 (13.9)
(b) Threshold ($m\ s^{-1}$)					
FF > 6.2	2720 (2570)	0.186 (0.179)	0.003 (0.010)	0.054 (0.045)	27.3 (19.1)
FF > 10.8	2720 (2570)	0.210 (0.214)	0.008 (0.027)	0.065 (0.056)	27.0 (13.8)
FF > 16.0	2720 (2570)	0.048 (0.051)	0.004 (0.011)	0.010 (0.007)	11.5 (−13.6)
(c) Period					
1	1632 (1554)	0.149 (0.147)	0.005 (0.009)	0.078 (0.072)	49.0 (42.9)
2	1632 (1554)	0.150 (0.145)	0.009 (0.017)	0.052 (0.042)	29.0 (17.0)
3	1632 (1554)	0.146 (0.144)	0.007 (0.018)	0.041 (0.037)	23.3 (12.9)
4	1632 (1554)	0.147 (0.144)	0.011 (0.028)	0.031 (0.028)	13.7 (−0.1)
5	1632 (1554)	0.147 (0.144)	0.010 (0.027)	0.027 (0.021)	11.8 (−4.2)
(d) Forecaster					
A	2760 (2430)	0.149 (0.141)	0.012 (0.022)	0.048 (0.034)	24.1 (8.5)
B	2010 (2355)	0.152 (0.146)	0.004 (0.022)	0.045 (0.033)	26.5 (7.2)
C	2415 (1905)	0.146 (0.143)	0.007 (0.019)	0.047 (0.046)	26.9 (19.3)
D	735 (825)	0.142 (0.163)	0.016 (0.010)	0.046 (0.069)	21.2 (36.1)

labelling task (presumably as a result of feedback and experience). Resolution, on the other hand, is more closely related to the state of the art of weather forecasting and reflects a forecaster's ability to "sort" the set of occasions into subsets associated with different relative frequencies of occurrence of the events [the use of the labelling and sorting terminology in this context was introduced by Sanders (1963)]. Since improvements in the state of the art of weather forecasting occur relatively infrequently and generally are quite modest, it is not surprising that only small changes occurred in the resolution of the experimental probability forecasts from year 1 to year 2. Significant improvements of this type must await increased understanding (on the part of the forecasters) of the behavior and evolution of the relevant weather systems or the availability of better guidance information based on numerical and/or statistical models.

TABLE 3. Partition of average ranked probability score and skill score based on average ranked probability score for experimental probability forecasts of visibility (VV) for second year: (a) overall; (b) by threshold; (c) by forecast period; and (d) by forecaster. Corresponding figures for first year are given in parentheses.

Stratification	Number of probabilities	Partition of ranked probability score: Variance of observations	Reliability	Resolution	Skill score (%)
(a) None	8160 (7770)	0.138 (0.141)	0.007 (0.018)	0.038 (0.035)	22.4 (12.4)
(b) Threshold (km)					
VV < 1	2720 (2570)	0.026 (0.031)	0.008 (0.011)	0.003 (0.006)	−20.9 (−16.0)
VV < 4	2720 (2570)	0.142 (0.148)	0.004 (0.022)	0.041 (0.038)	26.4 (10.6)
VV < 10	2720 (2570)	0.247 (0.246)	0.008 (0.021)	0.069 (0.063)	24.6 (17.0)
(c) Period					
1	1632 (1554)	0.139 (0.140)	0.005 (0.012)	0.073 (0.064)	48.4 (37.4)
2	1632 (1554)	0.134 (0.139)	0.014 (0.023)	0.037 (0.042)	17.4 (13.8)
3	1632 (1554)	0.142 (0.145)	0.012 (0.021)	0.032 (0.034)	14.8 (9.0)
4	1632 (1554)	0.133 (0.140)	0.009 (0.026)	0.030 (0.031)	15.7 (3.6)
5	1632 (1554)	0.141 (0.143)	0.010 (0.024)	0.030 (0.022)	14.3 (−2.0)
(d) Forecaster					
A	2760 (2430)	0.133 (0.141)	0.007 (0.018)	0.047 (0.033)	30.1 (10.8)
B	2010 (2355)	0.135 (0.136)	0.006 (0.034)	0.027 (0.034)	16.3 (−0.3)
C	2415 (1905)	0.139 (0.140)	0.009 (0.011)	0.032 (0.043)	16.6 (22.7)
D	735 (825)	0.155 (0.157)	0.019 (0.015)	0.057 (0.054)	24.4 (24.7)

TABLE 4. Partition of average ranked probability score and skill score based on average ranked probability score for experimental probability forecasts of precipitation (RR) for second year: (a) overall; (b) by threshold; (c) by forecast period; and (d) by forecaster. Corresponding figures for first year given in parentheses.

Stratification	Number of probabilities	Partition of ranked probability score: Variance of observations	Reliability	Resolution	Skill score (%)
(a) None	5440 (5140)	0.122 (0.143)	0.017 (0.056)	0.018 (0.021)	0.5 (−24.7)
(b) Threshold (mm)					
RR > 0.2	2720 (2570)	0.159 (0.186)	0.017 (0.057)	0.028 (0.032)	6.8 (−13.6)
RR > 1.4	2720 (2570)	0.085 (0.101)	0.017 (0.055)	0.008 (0.010)	−11.2 (−45.0)
(c) Period					
1	1088 (1028)	0.126 (0.135)	0.006 (0.034)	0.043 (0.051)	29.4 (12.9)
2	1088 (1028)	0.114 (0.142)	0.021 (0.054)	0.017 (0.023)	−3.5 (−21.6)
3	1088 (1028)	0.128 (0.139)	0.022 (0.077)	0.016 (0.016)	−4.9 (−43.8)
4	1088 (1028)	0.114 (0.149)	0.036 (0.067)	0.014 (0.016)	−19.6 (−34.2)
5	1088 (1028)	0.127 (0.150)	0.018 (0.067)	0.016 (0.015)	−1.3 (−34.5)
(d) Forecaster					
A	1840 (1620)	0.127 (0.135)	0.018 (0.044)	0.021 (0.022)	2.1 (−16.7)
B	1340 (1570)	0.119 (0.141)	0.027 (0.095)	0.025 (0.016)	−2.0 (−56.3)
C	1610 (1270)	0.114 (0.159)	0.010 (0.043)	0.016 (0.031)	4.5 (−8.0)
D	490 (550)	0.141 (0.142)	0.041 (0.042)	0.022 (0.036)	−13.8 (−4.0)

Substantial improvements in the quality of the experimental subjective probability forecasts formulated at Zierikzee were made from year 1 to year 2. What are the prospects for further improvements in the future? First, while additional feedback and experience can be expected to reduce still further the forecasters' tendency toward overforecasting, these improvements are unlikely to be as large as the improvements realized

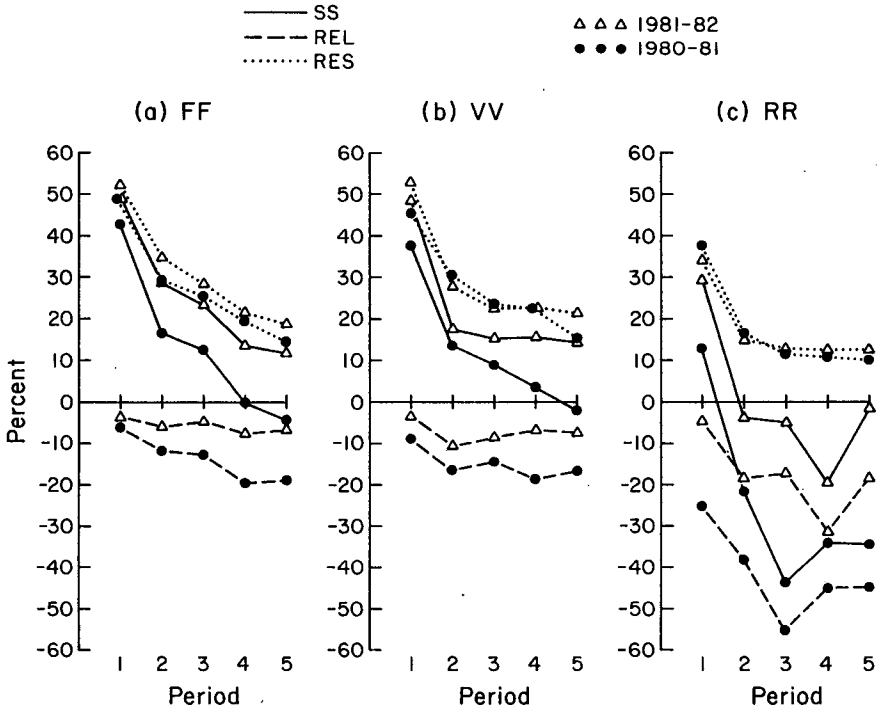

FIG. 6. Skill scores (SS) and skill score contributions of reliability (REL) and resolution (RES) terms, all in percent, for experimental probability forecasts for each element as function of period (all thresholds and forecasters combined). (a) Wind speed (FF), (b) visibility (VV), (c) precipitation (RR).

from year 1 to year 2. That is, a substantial fraction of the total possible improvement in reliability has already been captured. However, even modest reductions in the reliability term (in the three-term partition) could lead to noticeable improvements in some skill scores (e.g., from negative to positive values). With regard to advances in the state of the art of forecasting these wind speed, visibility, and precipitation events (i.e., improvements in resolution), such advances will require increased understanding of the dominant weather systems and the availability of guidance information as noted above. With regard to the latter, the availability of objective guidance forecasts for these events could lead to increased resolution and improved skill scores.

In conclusion, the results of the experimental probability forecasting program at Zierikzee presented in DM and in this paper are encouraging and indicate that forecasters can improve their probability assessments on the basis of feedback and experience. Specifically, the amount of overforecasting was reduced substantially from year 1 to year 2. As a result, marked improvements were observed in the skill scores. Moreover, in judging the results of this experiment, several factors must be kept in mind. First, it should be noted that the wind speed and visibility forecasts were of primary concern to the forecasters at Zierikzee, because of their operational significance, and this factor may explain in part the difference in quality between these forecasts and the precipitation forecasts (the latter clearly were less successful). Second, no specific guidance forecasts—probabilistic or nonprobabilistic—for these events were available to the forecasters during the experimental period. Third, sample climatological probabilities have been used as the standard of reference in computing skill scores in this study, and it is well known that the use of such sample probabilities underestimates the skill of the forecasts (e.g., see Murphy and Daan, 1984). Thus, as noted in DM, the results of the subjective probability forecasting experiment at Zierikzee—when considered in conjunction with other subjective probability forecasting programs in the United States and elsewhere (e.g., Murphy, 1981; Murphy and Winkler, 1982; Winkler and Murphy, 1979)—lend further support to the belief that experienced weather forecasters can quantify the uncertainty inherent in forecasts of a wide variety of elements or events in a reliable and skillful manner. Moreover, the results presented in this paper indicate that tendencies toward overforecasting (or underforecasting) on the part of such forecasters can be reduced by providing the forecasters with detailed feedback concerning their performance.

Acknowledgments. The participation of the first author (A.H.M.) in this study was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grants ATM-8004680 and ATM-8209713. The authors gratefully acknowledge the assistance and cooperation of the forecasters from the Royal Netherlands Meteorological Institute who are involved in the experimental probability forecasting program at Zierikzee, The Netherlands.

REFERENCES

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

Daan, H., 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.

Murphy, A. H., 1970: The ranked probability score and the probability score: A comparison. *Mon. Wea. Rev.*, **98**, 917–924.

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

——, 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.

——, 1981: Subjective quantification of uncertainty in weather forecasts in the United States. *Meteor. Rundsch.*, **34**, 65–77.

——, and R. L. Winkler, 1982: Subjective probabilistic tornado forecasts: Some experimental results. *Mon. Wea. Rev.*, **110**, 1288–1297.

——, and H. Daan, 1984: Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, Eds. Westview Press (in press).

——, W.-R. Hsu and R. L. Winkler, 1982: Subjective probabilistic quantitative precipitation forecasts: Some experimental results. *Preprints Ninth Conf. Weather Forecasting and Analysis*, Seattle, Amer. Meteor. Soc., 94–100.

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

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

Wallsten, T. S., and D. V. Budescu, 1983: Encoding subjective probabilities: A psychological and psychometric review. *Management Sci.*, **29**, 151–173.

Winkler, R. L., and A. H. Murphy, 1979: The use of probabilities in forecasts of maximum and minimum temperatures. *Meteor. Mag.*, **108**, 317–329.