Use of an Analogue Procedure to Formulate Objective Probabilistic Temperature Forecasts in The Netherlands

SEIJO KRUIZINGA

Royal Netherlands Meteorological Institute, De Bilt, The Netherlands

ALLAN H. MURPHY

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

(Manuscript received 23 February 1983, in final form 25 July 1983)

ABSTRACT

The purpose of this paper is to describe some results of a study in which an analogue procedure developed in The Netherlands is used to formulate objective probabilistic temperature forecasts on an experimental basis. As currently employed, the procedure routinely provides forecasters at the Royal Netherlands Meteorological Institute with guidance information that summarizes, for days 1 through 6, the weather conditions associated with the best thirty analogues of the corresponding forecast situation. In the work reported here, the empirical frequency distribution of maximum temperature corresponding to these thirty analogues is used to generate both categorical and probabilistic forecasts of this element. Attention is focused on three types of probabilistic forecasts of maximum temperature: 1) a discrete distribution for five temperature classes; 2) a variable-width credible interval; and 3) a fixed-width credible interval.

Results of the experiment indicate that all three types of probabilistic temperature forecasts are quite reliable, in the sense that the forecast probabilities correspond closely to the relative frequencies of observed temperatures associated with these classes/intervals. Moreover, the forecasts generally are more accurate and precise, according to several different measures of performance, than forecasts based on standards of reference such as climatology and persistence. Thus, these experimental objective probability forecasts usually exhibit positive skill. As expected, the level of skill decreases markedly from day 1 to day 6 for all three types of probabilistic forecasts. Evaluation of two types of categorical forecasts—the median and mean temperatures derived from the empirical frequency distribution—reveals similar results.

The implications of the results of this study for operational temperature forecasting are discussed briefly, and some possible refinements and/or improvements in objective probabilistic temperature forecasting are described.

1. Introduction

A strong case has been made recently for formulating and expressing day-to-day forecasts of surface temperature in probabilistic terms (Murphy and Winkler, 1979). This case is based primarily on two important considerations: 1) all weather forecasts are inherently uncertain and 2) the value of probabilistic forecasts generally exceeds the value of forecasts expressed in traditional categorical (i.e., nonprobabilistic) terms. Of course, the uncertainty in temperature forecasts can be quantified subjectively, and the results of several experiments reveal that experienced weather forecasters can quantify the uncertainty in their maximum and minimum temperature forecasts in a reliable and skillful manner (e.g., see Murphy and Winkler, 1974; Winkler and Murphy, 1979). Nevertheless, it is also desirable to investigate the possibility of using objective procedures to prepare probabilistic temperature forecasts. Such objective probabilistic forecasts should represent a valuable source of information for weather forecasters who must formulate the official temperature forecasts, especially when the latter are to be expressed in probabilistic terms. Moreover, in some circumstances, it may be necessary and appropriate to transmit objective probabilistic temperature forecasts directly to specific users and/or the general public.

In recent years, objective procedures based on the perfect prognostic (PP) and model output statistics (MOS) approaches have been used to produce surface temperature forecasts on an operational and/or experimental basis in several countries (e.g., Carter *et al.*, 1979; Conte *et al.*, 1980; Duvernet and Rousseau, 1980; Francis *et al.*, 1982; Wilson and Yacowar, 1980; Woodcock, 1980). Moreover, the forecasts provided by these procedures have demonstrated considerable skill. However, objective temperature forecasts generally have been formulated in categorical terms; that is, the uncertainty inherent in the forecasts has not been specified. Thus, little if any experience has been gained in objective probabilistic temperature forecasting. The purpose of this paper is to describe some results of a study in which an analogue procedure developed in The Netherlands is used to formulate objective probabilistic temperature forecasts on an experimental basis.

The analogue procedure and its operational use at the Royal Netherlands Meteorological Institute (KNMI) are described briefly in Section 2. Categorical and probabilistic temperature forecasts formulated and evaluated in this study are defined in Section 3. A short discussion of the methods used to evaluate the results of the temperature forecasting experiment is included in Section 4. Section 5 describes the results of the study, with emphasis on the reliability and skill of the different types of objective probabilistic temperature forecasts. Section 6 contains a discussion and conclusion.

2. The KNMI analogue technique

In the first half of this century, so-called "classical" analogue techniques were used frequently to obtain objective forecasts of meteorological fields and weather elements at specific locations (e.g., see Baur, 1951; Namias, 1951). However, use of the classical approach to forecast several days in advance generally would require that the analogues be based on information from a very large (e.g., hemispheric) grid of points. The availability of numerical weather prediction (NWP) models provides a means of reducing the size of the area over which analogues must be considered, since these NWP models are now quite capable of forecasting the development and evolution of weather systems on larger scales. This approach, in which analogues are used in conjunction with the output of NWP models to "translate" large-scale predictions into local surface weather forecasts, has been employed recently by Balzer (1976), Wilson and Yacowar (1980), Woodcock (1980), and Yacowar (1975) among others.

An analogue technique first was used to interpret the output of NWP models at KNMI in 1970 when this approach was employed in conjunction with the 72-hour 500 mb forecasts produced by the National Meteorological Center in the United States. Initially, the selection of analogues was based on a subjective comparison of the predicted 500 mb field with a series of historical 500 mb maps, and the weather associated with the best analogue was used as guidance. In 1975, this subjective system was replaced by an objective procedure and the operational guidance system evolved from considering the weather associated with the best single analogue to using a statistical summary of the conditions corresponding to the best thirty analogues. In this section, we provide a short description of the objective procedure currently employed to select the analogues and give an example of the information provided as guidance to KNMI forecasters.

The selection of analogues is presently based on the 500 mb forecasts produced by the European Center for Medium Range Weather Forecasts (ECMWF) in Reading, United Kingdom. Analogues are chosen for six equally-spaced lead times from 24 hours to 144 hours (subsequently denoted by day 1 to day 6) based on the corresponding ECMWF prognostic charts, and these forecasts are all valid at 1200 GMT. In selecting the analogues, only the 500 mb heights at the 58 grid points indicated in Fig. 1 was considered. The historical data set scanned for suitable analogues consists of 500 mb heights at these same 58 grid points each day at 0000 GMT for the period 1 January 1949–31 December 1976. Before the degree of similarity between the predicted 500 mb height field and the 500 mb height field associated with a potential analogue is computed, a condition involving the differences in dates between the two charts must be satisfied. Specifically, only those historical 500 mb height fields whose dates differ from the date of the prognostic 500 mb height field by, at most, twenty days are considered as potential analogues.

For each analogue that satisfies the criterion concerning differences in dates, a similarity measure S is computed between the two fields. This measure is defined as

$$S = \sum_{n=1}^{58} w_n[(F_n - \bar{F}) - (A_n - \bar{A})]^2, \qquad (1)$$

where F_n and A_n denote the values of the 500 mb heights at the nth grid point of the ECMWF forecast and potential analogue respectively, $\bar{F}$ and $\bar{A}$ represent the weighted averages of the respective fields, and w_n denotes the weight attached to the nth grid point. With regard to the latter, the weights attached to the grid points are indicated in Fig. 1. (These weights were chosen as a result of a small pilot study in which alternative procedures and weights were considered. Different sets of weights appeared to have relatively little effect on the set of analogues actually selected. The weights used here led to the selection of sets of analogues that corresponded more closely to the sets of analogues chosen subjectively by forecasters than analogues associated with other weights.) These weights also are used in the computation of the weighted averages; that is,

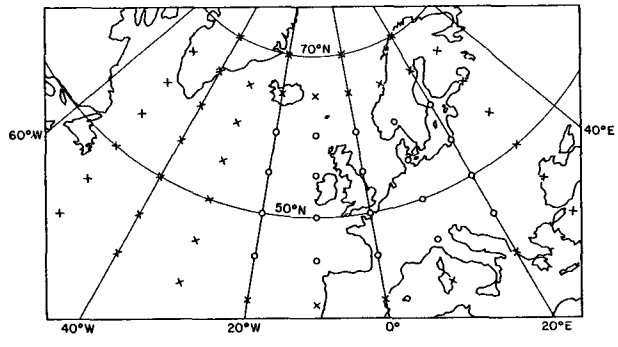

FIG. 1. Array of 58 grid points at which 500 mb heights are considered in selection of thirty best analogues. Weights assigned to heights at gridpoints are defined by symbols: × = 1 and ○ = 3.

$$\bar{F} = \frac{\sum_{n=1}^{58} w_n F_n}{\sum_{n=1}^{58} w_n}, \qquad (2)$$

and $\bar{A}$ is defined in a similar manner. The similarity measure S in Eq. (1) is zero in the case of a perfect match, and smaller values of S indicate better analogues. In the scan of the historical data set, the dates and values of the similarity measure for the thirty best analogues (thirty lowest values of S) are retained.

After the thirty best analogues have been identified, certain statistics summarizing the surface weather conditions associated with these analogues are presented to KNMI forecasters as guidance. [All weather information is taken from the 24-hour period following the validation time of the analogues. Furthermore, the difference in validation times of the NWP forecasts (1200 GMT) and the analogues (0000 GMT) is assumed to be of minor importance and therefore is ignored.] An example of the guidance provided to the forecasters on 1 December 1982 for a 48-h forecast is contained in Table 1. Note that guidance information is summarized in the form of empirical relative frequencies for sunshine at De Bilt (four categories), precipitation amount at De Bilt (four categories), areal coverage of precipitation in The Netherlands (four categories), and maximum and minimum temperatures at De Bilt (thirteen categories each). The average maximum and minimum temperatures and the standard deviations of these temperatures also are indicated. Finally, dates of the thirty analogues are listed in chronological order.

The example presented in Table 1 suggests that samples of analogues selected by this procedure may tend to cluster into small groups by date. When this clustering occurs, it appears to imply that the weather conditions associated with such analogues are not independent in a statistical sense, leading to a reduction in the effective sample size. However, persistence in weather conditions is due in large measure to persistence in the atmospheric circulation patterns. Thus samples of weather conditions associated with specific patterns may provide largely independent information concerning the relevant distributions, *given a specific pattern.*

3. Nature and type of forecasts

The forecasts of interest in this study are based upon the empirical (relative) frequency distributions of temperature associated with the thirty best analogues on the relevant occasions. All temperatures are expressed as deviations from their respective pentad normals. Let $T_1, \ldots, T_{30}$ denote the thirty temperatures associated with the analogues on a particular occasion and let $T^{(1)}, \ldots, T^{(30)}$ represent the ordered set of these temperatures (from lowest to highest). Finally, let T_0 represent the observed temperature on that occasion. In this section we define the categorical and probabilistic temperature forecasts of interest using this notation.

Two categorical, or point, temperature forecasts are formulated on the basis of each empirical frequency

TABLE 1. Example of analogue-based guidance provided to KNMI forecasters on 1 December 1982 for a 48-h forecast.

FREQUENTIE VERDELING DE BILT OP BASIS VAN DE 48 UUR PROG GELDIG VOOR DONDERDAG 2 DECEMBER 1982

	1	2	3	4	ONTBR
SS	53	23	13	10	0
RRDB	90	0	3	7	0
RRLAND	63	30	3	3	0

	<−5	−5	−4	−3	−2	−1	0	1	2	3	4	5	>5	GEM	STD
TN	7	3	3	7	23	13	13	10	3	0	7	3	7	−0.5	3.3
TX	10	10	10	10	13	10	0	7	17	13	0	0	0	−1.5	3.2

531207	531208	531209	531210	531211	531212	531213	531221	561121	561122
571118	571119	571120	581117	581118	581119	581120	581121	581124	631128
631129	681204	681205	701212	711206	711218	721220	721221	721222	781112

Key

SS	Sunshine duration relative to maximum possible duration (4 classes): (1) 0%; (2) 1–29%; (3) 30–59%; (4) 60–100%.
RRDB	Precipitation at de Bilt, 1800–1800 GMT (4 classes): (1) <0.3 mm; (2) 0.3–1.4 mm; (3) 1.5–4.4 mm; (4) ≥4.5 mm.
RRLAND	Areal coverage of precipitation, in terms of number of stations out of 10 key stations (4 classes): (1) 0 stations; (2) 1–2 stations; (3) 3–7 stations; (4) 8–10 stations.
ONTBR	Number of missing observations.
TN	Minimum temperature at de Bilt, 1800–0600 GMT, in terms of deviation from pentad normal (°C).
TX	Maximum temperature at de Bilt, 0600–1800 GMT, in terms of deviation from pentad normal (°C).
GEM	Average minimum and maximum temperatures (°C).
STD	Standard deviation of minimum and maximum temperatures (°C).

List of 30 best analogues in chronological order by year, month and day.

distribution; namely, the mean and the median. The mean temperature forecast is denoted by $\bar{T}$, where

$$\bar{T} = I\left[\left(\frac{1}{30}\right)\sum_{i=1}^{30} T_i\right], \quad (3)$$

in which $I[x]$ is the nearest integer to x. The median temperature forecast is denoted by T_M, where

$$T_M = I\left[\frac{1}{2}\left(T^{(15)} + T^{(16)}\right)\right]. \quad (4)$$

In addition to the point forecasts $\bar{T}$ and T_M, several other categorical forecasts are considered as standards of reference and these other point forecasts are defined in Section 4.

Three types of probability forecasts are derived from the basic empirical frequency distribution: 1) a discrete probability distribution for a set of five temperature classes, 2) a variable-width credible interval, and 3) a fixed-width credible interval. The set of five mutually exclusive and collectively exhaustive classes of temperature values for the discrete probability forecast is specified in Table 2. These classes are defined with respect to the pentad normal. In this case, a forecast consists of the relative frequencies assigned to the five classes by the analogue technique.

A credible interval temperature forecast is an interval of temperature values accompanied by a probability that indicates the likelihood that the observed temperature will fall in the interval (e.g., see Murphy and Winkler, 1974). The variable-width (fixed-probability) credible interval forecast of interest here is defined as the temperature interval from $T^{(6)}$ to $T^{(25)}$ inclusive. Thus, this interval is a central credible interval (i.e., it is centered at the median T_M in terms of probability), and it contains approximately 66.7% (i.e., two-thirds) of the cases in the empirical probability distribution. [Note: In reality, $T^{(6)}$ and $T^{(25)}$ are estimates of the 0.194th ($=6/31$) and 0.806th ($=25/31$) quantiles, respectively, of the cumulative distribution, leading to a variable-width central credible interval with probability 0.612. However, the discrete nature of this distribution means that the interval generally will contain more than 61.2% of the thirty temperatures (e.g., see Table 1). Simulations conducted by a referee and based on a Gaussian distribution suggest that approximately 70–75% of the temperatures should fall in the interval. Thus, the interval from $T^{(6)}$ to $T^{(25)}$ would be expected to include somewhat more than two-thirds of the observed temperatures.]

TABLE 2. Classes of temperature values used to define events for discrete probability forecasts.

Class number	Interval of temperature values (°C)	Class name (symbol)
1	$T_0 \leq -5$	Much below (MB)
2	$-4 \leq T_0 \leq -2$	Below (B)
3	$-1 \leq T_0 \leq 1$	Normal (N)
4	$2 \leq T_0 \leq 4$	Above (A)
5	$5 \leq T_0$	Much above (MA)

The fixed-width (variable-probability) credible interval forecast is based on the temperature interval from $\bar{T} - 2$°C to $\bar{T} + 2$°C inclusive. Thus, the width of this interval is 5°C, and the probability assigned to the interval on a particular occasion is determined by adding the relative frequencies associated with these five temperature values in the empirical distribution on that occasion. It should be noted that the fixed-width intervals are central intervals in terms of width but not in terms of probability. Standards of reference for these three types of probability forecasts will be defined in Section 4.

4. Methods of evaluation

In this section we describe briefly the methods used to evaluate the categorical and probabilistic temperature forecasts produced by the analogue technique. Since the quality of the probability forecasts is of primary concern, particular emphasis is placed on methods of evaluating these forecasts. The attributes of the forecasts of special interest are accuracy, reliability, and skill. Forecasts to be considered as standards of reference also are identified here.

The basic measure of the accuracy of categorical forecasts used in this paper is the mean absolute error. This measure is employed to compare the quality of the two types of point forecasts (mean $\bar{T}$ and median T_M) with the performance of three standards of reference; namely, a climatological forecast (T_c), a persistence forecast (T_p), and a regression forecast (T_r) based on a linear combination of persistence and climatology. The climatological forecast is a (constant) forecast of the pentad normal. A persistence forecast for day k ($k = 1, \ldots, 6$) consists of the observed temperature (T_0) k days before the day in question (i.e., on the initial day). Finally, the regression forecast is the integer closest to an expression in which the persistence temperature value is multiplied by the appropriate first-order autocorrelation coefficient (Daan, 1980).

The measures used to evaluate the probabilistic forecasts depend upon the nature of these forecasts. In the case of the discrete probability forecasts defined on the (fixed) set of five classes of temperature values, two common scoring rules for probabilistic forecasts—the Brier score (BS) (Brier, 1950) and the ranked probability score (RPS) (Epstein, 1969; Murphy, 1971)—are employed. Actually, the RPS appears to be more appropriate in the context of temperature forecasting, since this measure (unlike the BS) takes the ordinal nature of the variable into account. Standard skill scores based on the BS and the RPS are used to compare the accuracy of these probability forecasts to the

accuracy of forecasts based upon climatological probabilities (e.g., see Daan and Murphy, 1982).

The reliability of the variable-width credible interval forecasts is evaluated by comparing the relative frequencies of observed temperatures below, in, and above the intervals with the forecast probabilities associated with these intervals (namely, 0.167, 0.667, and 0.167, respectively). The variable-width forecasts also are evaluated using a loss function $L = L(T^{(6)}, T^{(25)}, T_0)$, where

$$L = \begin{cases} 6(T^{(6)} - T_0) + (T^{(25)} - T^{(6)} + 1), \\ \qquad \text{if} \quad T_0 < T^{(6)} \\ T^{(25)} - T^{(6)} + 1, \quad \text{if} \quad T^{(6)} \leq T_0 \leq T^{(25)} \\ 6(T_0 - T^{(25)}) + (T^{(25)} - T^{(6)} + 1), \\ \qquad \text{if} \quad T_0 > T^{(25)} \end{cases} \tag{5}$$

(see Winkler, 1972; Winkler and Murphy, 1979). The average losses for a set of forecasts can be considered to represent the average "expenses" of an individual who makes decisions on the basis of the forecasts. These average losses reflect both the reliability and precision (i.e., width) of the relevant forecasts.

The fixed-width credible interval forecasts are evaluated by computing a score essentially equivalent to the Brier score. This score is the square of the difference between the probability assigned to the interval and unity when the observed temperature falls in the interval and simply the square of the probability when the observed temperature falls above or below the interval. The average score for the forecasts of interest is then compared with average scores for three types of forecasts (denoted here by R1, R2 and R3). The R1 forecasts are produced by assigning the fixed interval $[-2, +2]$ the appropriate monthly climatological probability. The interval $[\bar{T} - 2, \bar{T} + 2]$ is used in formulating the R2 forecasts, and the corresponding probability is determined by adding the monthly climatological probabilities associated with these five temperature values. The R3 forecasts also involve the use of the interval $[\bar{T} - 2, \bar{T} + 2]$, but in this case it is assigned a fixed probability for each lead time which is equal to the observed relative frequency of temperatures in the interval during the first year of the experimental period.

5. Results

As indicated in Section 3, both categorical and probabilistic forecasts can be derived from the distributions of temperatures associated with the sets of thirty best analogues. Although we are concerned primarily with the quality of the probabilistic temperature forecasts, Section 5a contains a short discussion of the accuracy of the objective categorical temperature forecasts. The reliability and skill of the probabilistic temperature forecasts are investigated in some detail in Section 5b. The results presented here are based on maximum temperature forecasts for De Bilt formulated during the two-year period from 1 December 1980 through 30 November 1982.

a. Categorical forecasts

The mean absolute error (MAE) of the average temperature forecasts ($\bar{T}$) is depicted in Fig. 2 as a function of lead time in days (curve A). The curve for the median temperature forecasts (T_M) is not included in the figure because it exhibits a very similar behavior to the curve for $\bar{T}$ (in fact, the former exceeds the latter by approximately 0.02°C at each lead time). As expected, the MAE for $\bar{T}$ increases as lead time increases, from slightly less than 2.0°C at day 1 to 2.5°C at day 6.

The MAE curves for the three standards of reference—namely, climatology (C), persistence (P), and regression (R)— also are shown in Fig. 2. Note that the analogue forecasts are more accurate than the forecasts produced by the reference procedures at all lead times, although the differences in accuracy between A and R and A and P are quite small on day 1. As expected, the accuracy of the persistence forecasts deteriorates rapidly as lead time increases, whereas the MAE for climatological forecasts is nearly constant. The skill of the $\bar{T}$ forecasts, measured in terms of percentage improvement of the MAE for these forecasts over the MAE for climatological forecasts, decreases from 35% on day 1 to 17% on day 6.

The numbered days in Fig. 2 represent the MAEs for subjective (maximum) temperature forecasts prepared by the Forecast Division at KNMI during the same two-year period. These forecasts are for days 1, 2, 3 and 4 although they are plotted at days 2, 3, 4

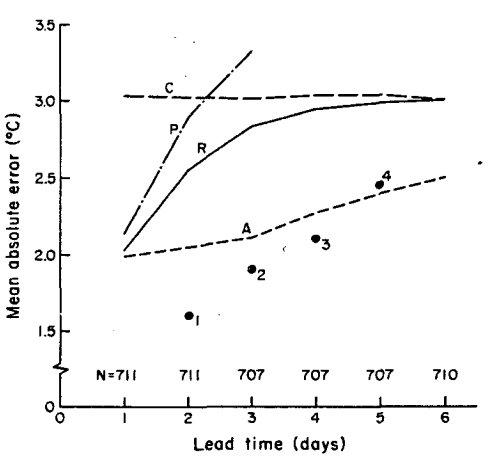

FIG. 2. Mean absolute errors (MAEs) (°C) as function of lead time for analogue-based mean temperature forecasts (curve A) and for forecasts based on climatology, persistence, and regression (curves C, P, and R, respectively). MAEs for subjective forecasts denoted by closed circles (see text for additional details).

and 5, respectively (the forecasters had access to the analogues from "yesterday" at the time of preparation of their forecasts). Sample sizes for the subjective and objective forecasts are similar except for day 4 (5) when approximately 10% of the subjective forecasts is missing. The subjective forecasts improved upon the objective analogue forecasts at all lead times, with the magnitude of improvement decreasing as lead time increased. The analogue forecasts are regarded as especially useful guidance for day 2 (3) and beyond.

Statistical interpretation of the output of numerical models using an analogue technique is necessarily based on a perfect prognostic approach (Klein, 1969). Thus, the forecasts may exhibit a bias as a result of numerical model bias. In this case, however, calculation of the mean (arithmetic) error in the temperature forecasts reveals that the bias varied from 0.12°C on day 1 to −0.15°C on day 6. Thus, the bias in the forecasts appears to be quite small.

Another issue of concern is the extent to which the use of the thirty best analogues in formulating (categorical) temperature forecasts leads to "conservative" forecasts. It was found that, over all lead times, the standard deviation of the mean temperature forecasts $\bar{T}$ is approximately 65% of the standard deviation of the observed temperatures. Thus, the variability of the (categorical) temperature forecasts is substantially smaller than that of the observations.

In the foregoing discussion, we have been concerned with the performance of the mean (and median) temperature derived from the distribution of temperatures associated with the thirty best analogues. In addition to measures of location (such as the mean), measures of the variability (or spread) of the distribution are of interest. In fact, such measures provide important insight into whether or not the distribution is able to "characterize" the uncertainty in the forecasts. In this regard, we have stratified the sample of temperature forecasts for each lead time into subsamples according to the standard deviation (*STD*) of the distributions and then computed the mean absolute error of the forecasts in these subsamples. The subsamples correspond to the following ranges of *STD*: $STD < 2$°C, $2°C \leq STD < 3°C$, $3°C \leq STD < 4°C$, and $STD \geq 4°C$. The results are presented in Fig. 3 (the number of cases associated with the subsamples represent approximately 19, 45, 27 and 9% of the total data set irrespective of lead time). Clearly, lower standard deviations are associated with more accurate forecasts (i.e., smaller MAEs) at all lead times. This result strongly suggests that the temperature distributions associated with the analogues contain at least some information concerning the uncertainty in the forecasts. Examination of the slopes of the respective curves indicates that a stronger relationship exists between accuracy and lead time for the forecasts with low STDs, suggesting that the quality of the numerical model is more important for this subsample.

b. Probability forecasts

In Section 3 we described three types of probability forecasts that are derived from the basic empirical frequency distribution of maximum temperatures associated with the thirty best analogues; namely, 1) a discrete probability distribution for a set of five temperature classes (see Table 2); 2) a variable-width credible interval; and 3) a fixed-width credible interval. The three types of forecasts are considered in this order and, for convenience, are sometimes denoted here by C5, VW and FW, respectively.

1) FIVE-CLASS DISCRETE PROBABILITY FORECASTS (C5)

The reliability of the C5 forecasts is depicted in the form of reliability diagrams in Fig. 4. To conserve space, only the diagrams for days 2 and 5 are presented (the diagrams for other lead times exhibit similar behavior, and the day 2 and day 5 results were selected to represent short-range and medium-range forecasts, respectively). In plotting the points in these diagrams, we divided the range of probability values into ten intervals of equal width (i.e., 0–10%, 10–20%, . . . , 90–100%), and the probabilities constituting the forecasts were assigned to subsets corresponding to the intervals (regardless of the class with which the probability was associated). Then the relative frequency of occurrence of temperatures in the respective classes was determined for each subset; this observed relative frequency was plotted at the midpoint of the corresponding (probability) interval.

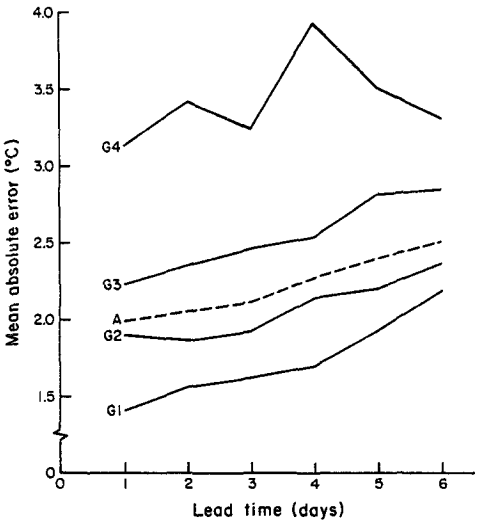

FIG. 3. Mean absolute errors (°C) as function of lead time for analogue-based temperature forecasts for four ranges of standard deviations (*STD*s) of basic forecast distributions: (1) G1: $STD < 2$°C, (2) G2: $2°C \leq STD < 3°C$, (3) G3: $3°C \leq STD < 4°C$, (4) G4: $STD \geq 4°C$. Dashed curve represents overall MAE (same as curve A in Fig. 2).

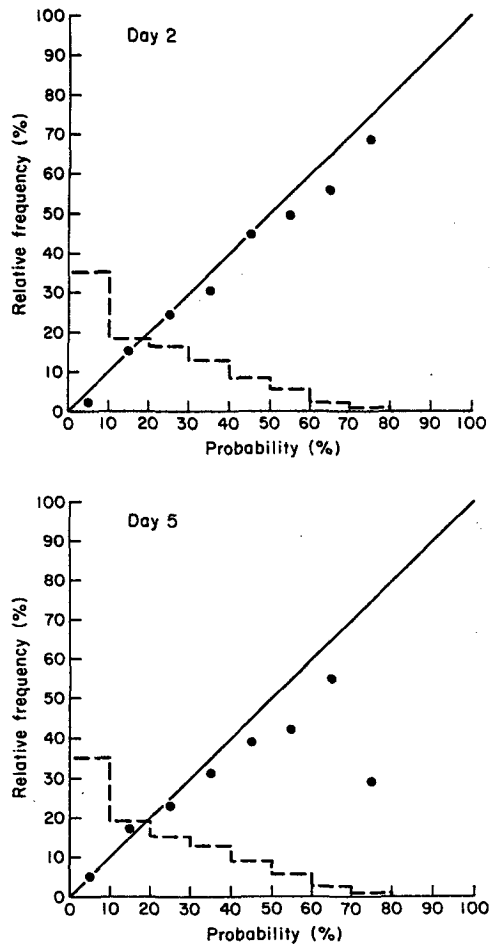

FIG. 4. Reliability diagrams for five-class discrete probability forecast (C5) for days 2 and 5. Histogram insets represent frequency of use distributions of probability values.

The curves in Fig. 4 indicate that the C5 forecasts are quite reliable, in the sense that these curves generally correspond closely to the 45° line representing perfect reliability. A slight tendency exists for the forecast probabilities to exceed the relative frequencies for larger probability values. The insets in the form of histograms in this figure represent the relative frequency of use of probability values in the various intervals. These distributions indicate that probabilities less than or equal to 50% were used on approximately 90% of the forecasting occasions.

The overall bias of the C5 forecasts by class is presented as a function of lead time in Table 3. Bias is defined as the difference between the average forecast probability and the overall observed relative frequency. A positive bias indicates that the average probability is too high, whereas a negative bias indicates that the average probability is too low. The results reveal some positive and negative biases for the normal, above normal, and much above normal classes. These biases may be due to differences in climatology between the historical 31-year period from which the analogues are selected and the two-year period of this study.

TABLE 3. Overall bias of five-class discrete probability forecasts (C5) by class.

Lead time (day)	Number of forecasts	Overall bias (%)				
		Much below	Below	Normal	Above	Much above
1	711	0	0	−2	4	−2
2	711	0	0	−2	4	−2
3	707	0	0	−3	4	2
4	707	0	0	−2	3	−2
5	707	1	0	−2	5	−2
6	710	1	1	−3	3	−3

The accuracy and skill of the C5 forecasts are evaluated in terms of the Brier score, ranked probability score, and the corresponding skill scores in Table 4. These scores reveal that the C5 forecasts possess positive skill for all lead times, with accuracy and skill decreasing as lead time increases (as expected). The differences between the two skill scores relate to the fact that the ranked probability score, unlike the Brier score, is sensitive to distance. Since the ranked probability score possesses this property, it takes the ordinal nature of a variable such as temperature into account in the evaluation process and gives "partial credit" for probabilities assigned to classes in the vicinity of the observed class. Thus, the fact that the skill scores based on the ranked probability score are higher than the skill scores based on the Brier score suggests that the C5 forecasts are more skillful than a "traditional" evaluation (in terms of the Brier score) would indicate.

2) VARIABLE-WIDTH CREDIBLE INTERVAL FORECASTS (VW)

The reliability of the VW forecasts is indicated in Table 5. Here, the relative frequencies of observed temperatures below, in, and above the variable-width intervals are compared with the probabilities corre-

TABLE 4. Brier score, ranked probability score and corresponding skill scores for five-class discrete probability forecasts (C5).

Lead time (day)	Number of forecasts	Brier score		Ranked probability score	
		Average score	Skill score (%)	Average score	Skill score (%)
1	711	0.652	15	0.449	32
2	711	0.657	14	0.449	32
3	707	0.671	12	0.468	29
4	707	0.688	10	0.504	24
5	707	0.715	7	0.530	21
6	710	0.738	3	0.556	16

TABLE 5. Relative frequency of observed temperatures below interval, in interval, and above interval, and average loss and skill score based on loss function for variable-width credible interval forecasts (VW).

Lead time (day)	Number of forecasts	Relative frequency of observed temperatures (%)			Loss function	
		Below interval	In interval	Above interval	Average loss	Skill score (%)
1	711	12.9	76.1	11.0	8.96	29
2	711	13.2	73.8	12.9	8.89	29
3	707	11.5	75.2	13.3	9.23	26
4	707	13.7	72.0	14.3	9.83	22
5	707	16.4	65.6	18.0	10.40	17
6	710	15.3	66.5	18.2	10.93	12

sponding to these subintervals (approximately 16.7, 66.7 and 16.7%, respectively). This comparison reveals that the relative frequency of temperatures in the interval exceeds the probability for days 1–4 (this result may be due in part to the fact that the distribution of temperatures is discrete—see note in Section 3), with the correspondence between probability and relative frequency being quite close for days 5 and 6. Relative frequencies below the interval exceed relative frequencies above the interval for days 1 and 2, whereas the opposite is true for days 3–6. While these differences are small, they are consistent with the biases in the mean temperature forecasts noted in Section 5a.

The precision of the VW forecasts can be measured in terms of the average width of the intervals. In this regard, the average width of the forecast intervals is approximately 5.8°C irrespective of lead time, whereas the average width of comparable climatological forecasts is 8.2°C. The loss function defined in Section 4 provides another measure of the precision of VW forecasts; the average losses for these forecasts are presented in Table 5. Such losses can be considered representative of the average "expenses" incurred by an individual who makes decisions on the basis of the forecasts. A corresponding skill score can be defined as the percentage improvement (i.e., reduction) in the expenses associated with the forecasts over the expenses associated with climatological forecasts. The results (Table 5) indicate that the skill score is positive for all lead times and that, as expected, the average losses increase and the skill score decreases as lead time increases.

For variable-width (fixed-probability) credible interval temperature forecasts, the relative frequency of occurrence of temperatures in the interval should be independent of the interval width. To investigate this property of the VW forecasts, the relative frequency of temperatures in the interval is plotted as a function of interval width for the day 2 and day 5 forecasts in Fig. 5. The dashed lines in the diagrams represent the best fitting straight lines to the respective sets of points. It is evident that a slight dependence exists, in that the relative frequency of temperatures in the interval tends to increase as the width of the interval increases. This result suggests that, in a relative sense, the narrower intervals are too narrow and the wider intervals are too wide.

Although the variable-width intervals are central credible intervals in terms of probability, they need not be symmetric about the median in terms of width. In this regard, the asymmetry of the VW forecasts was determined by computing the measure A, where $A = (T_M - T^{(6)}) - (T^{(25)} - T_M)$, in which T_M is the median temperature and $T^{(6)}$ and $T^{(25)}$ are the temperatures corresponding to the lower and upper endpoints of the variable-width credible interval. The frequency distribution of the measure A is presented in Table 6 for the day 2 and day 5 forecasts (the distributions of A for other lead times are similar). This table indicates that a substantial majority of the VW forecasts are symmetric or nearly symmetric (i.e., $|A| \leq 1$). However, approximately 25% of the VW forecasts are asymmetric, in that the absolute value of the measure A equals or exceeds two.

3) FIXED-WIDTH CREDIBLE INTERVAL FORECASTS (FW)

Reliability diagrams for the FW forecasts on days 2 and day 5 are presented in Fig. 6. These diagrams were prepared in a manner similar to the reliability diagrams in Fig. 4. The curves in Fig. 6 indicate that the FW forecasts are not as reliable as the C5 forecasts.

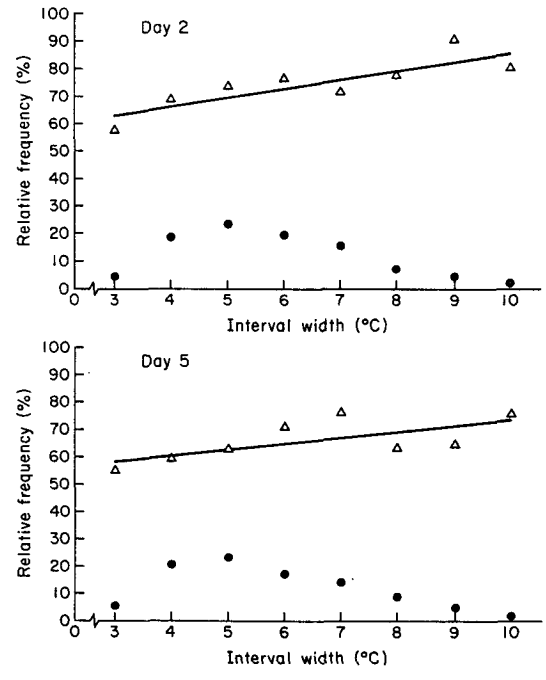

FIG. 5. Relative frequency of observed temperatures in variable-width credible intervals as function of interval width for days 2 and 5. Solid lines represent best fitting straight lines to sets of points. Frequency distributions of interval widths denoted by dots.

TABLE 6. Frequency distribution of asymmetry measure A for variable-width credible interval forecasts (VW) for day 2 and day 5.

Value of asymmetry measure A (°C)	Relative frequency (%)	
	Day 2	Day 5
−4	1.0	0.7
−3	2.2	2.4
−2	6.3	7.4
−1	20.1	20.2
0	31.5	30.4
1	24.3	23.3
2	10.7	11.2
3	2.4	3.8
4	1.3	0.3
5	0.1	0.3

In particular, relative frequencies tend to exceed forecast probabilities for lower probability values, whereas the opposite is true for higher probability values (i.e., the slopes of these curves are somewhat less than that of the 45° line representing perfect reliability). The frequency of use distributions (histogram insets) reveal that probabilities in the range from 50 to 90% were used on more than 75% of the forecasting occasions.

The overall bias of the FW forecasts is indicated in Table 7. These results reveal a systematic trend in the bias values from a negative value at day 1 to a positive value at day 6. Thus, a tendency exists for the average probability assigned to the FW interval to be too low for shorter lead times and too high for longer lead times.

The accuracy and skill of the FW forecasts also are described in Table 7. As expected, accuracy and skill decrease as lead time increases. The differences in the skill scores indicate that the FW forecasts are quite accurate relative to some standards of reference (e.g., R1), but are only marginally more accurate than other standards of reference (e.g., R3).

6. Discussion and conclusion

In this paper we have described some results of a study in which an analogue procedure developed in The Netherlands has been used to formulate objective maximum temperature forecasts for days 1–6 at De Bilt. The procedure is based on a perfect prognostic (PP) approach, utilizing the 500 mb prognostic charts produced by ECMWF. Both categorical and probabilistic forecasts were derived from the frequency distribution of maximum temperatures associated with the best thirty analogues. In this study, attention has been focused primarily on the reliability and skill of the probabilistic forecasts.

Two types of categorical temperature forecasts were considered—the mean and median of the frequency distribution—and the results of this study indicate that both types of forecasts are more accurate than reference forecasts based on climatology, persistence, or a combination of climatology and persistence. Moreover, the accuracy of these categorical forecasts, as measured by the mean absolute error, was found to be related to the variance of the frequency distribution of temperatures associated with the analogues. Specifically, the accuracy of the forecasts increased as the variance decreased, indicating that these frequency distributions contained information concerning the uncertainty in the forecasts.

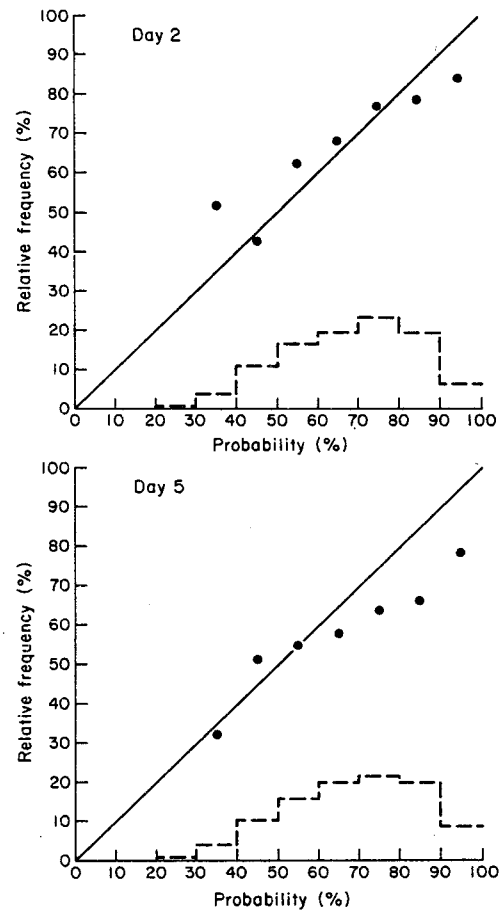

FIG. 6. As in Fig. 4 but for fixed-width credible interval forecasts (FW) for days 2 and 5.

TABLE 7. Bias, Brier score, and skill scores for fixed-width credible interval forecasts (FW).

Lead time (day)	Number of forecasts	Bias (%)	Brier score	Skill score (%)		
				S1	S2	S3
1	711	−5	0.202	28	19	3
2	711	−3	0.203	27	19	5
3	707	−1	0.218	21	13	1
4	707	1	0.217	19	14	5
5	707	7	0.241	9	4	−1
6	710	8	0.249	6	1	−4

Three types of probability forecasts were derived from each frequency distribution: 1) a discrete five-class forecast (C5); 2) a variable-width credible interval (VW); and 3) a fixed-width credible interval (FW). Both the C5 and the VW forecasts were quite reliable, in the sense that the forecast probabilities corresponded quite closely to the relative frequencies of observed temperatures in the classes/intervals. On the other hand, the FW forecasts exhibited a tendency for forecast probabilities to exceed relative frequencies for higher probability values. Thus, these forecasts were somewhat less reliable than the C5 and VW forecasts.

All three types of probability forecasts were found to be skillful when compared with simple reference forecasts based on climatology. However, the comparison of the FW forecasts with other standards of reference indicated that the skill of these forecasts resided in the placement of the interval on the temperature scale rather than in the day-to-day variability of the probability assigned to the interval. The results also revealed that the VW forecasts were more precise than forecasts based on climatological intervals, as reflected by the average widths of the intervals and by the average values of an appropriate loss function. Finally, approximately 25% of the VW forecasts were found to be asymmetric in terms of width.

The analogue forecasting procedure which has been employed here to formulate the objective categorical and probabilistic temperature forecasts is based on a perfect prognostic (pp) approach; that is, the ECMWF 500 mb prognostic charts employed in this procedure were treated as "perfect," in the sense that they were assumed to provide completely accurate descriptions of the height field at 500 mb at all future times. Then, for each day, the best thirty analogues of the predicted height field were sought in the historical file, and the frequency distribution of maximum temperatures associated with these analogues was assumed to define the basic probabilistic forecast of interest (see Section 2). Thus, as is well known, the PP approach does not take account of any systematic errors present in the ECMWF model forecasts (i.e., in the predicted 500 mb height field). The failure to consider such errors can lead to bias in the corresponding weather forecasts, especially at longer lead times. Moreover, Glahn (1983) has noted that probabilistic forecasts based on the PP approach may not tend toward climatological probabilities as lead time increases (a desirable property of any imperfect forecasting procedure), but rather may remain too "sharp" because the prognostic charts on which the forecasts are based are assumed to be perfect.

Some evidence of PP bias in the experimental probabilistic temperature forecasts can be seen in Tables 5 and 7, where differences between some forecast probabilities and corresponding relative frequencies tend toward smaller positive or larger negative values as lead time increases. However, these effects are not very large; in fact, the VW forecasts are more reliable at longer lead times than at shorter lead times (see Table 5). Thus, the use of the PP approach does not appear to have had any major adverse effect on the probabilistic forecasts. This conclusion is consistent with the experience of meteorologists at KNMI, who report that the analogue procedure is relatively insensitive to errors in numerical weather prediction models.

The analogue procedure employed in this study is simple to use (once a historical file of potential analogues has been created) and, as demonstrated here, it provides reliable and skillful temperature forecasts. However, several possible refinements in this procedure might be investigated in order to improve the quality of the resulting weather forecasts. For example, verification statistics from a large sample of ECMWF 500 mb prognostic charts could be used to estimate the bias in this model output, and these biases could be corrected before the historical file is searched for the best thirty analogues. Of course, such a correction procedure might be invalidated by a significant change in the ECMWF model in the same way that the model output statistics (MOS) approach can be affected adversely by substantial model changes. Alternatively, systematic errors in the temperature forecasts themselves could be corrected on the basis of verification statistics from a prior sample of such forecasts. Finally, the output of the analogue procedure, in one form or another, could serve as input (i.e., as predictors) to a regression model, with the latter being used to produce numerical–statistical forecasts of surface temperature (e.g., see Woodcock, 1980). Such a scheme could provide either categorical or probabilistic forecasts.

As noted in Section 1, the need to quantify the uncertainty in the daily surface temperature forecasts provided to both specific users and the general public has been emphasized by Murphy and Winkler (1979). While this uncertainty can be quantified subjectively by experienced weather forecasters, it undoubtedly would be desirable to provide such individuals with objective probabilistic temperature forecasts as guidance. Moreover, guidance forecasts expressed in probabilistic terms—such as the maximum temperature forecasts described in this paper—should be more useful to forecasters (and others) than guidance forecasts expressed in categorical terms, regardless of the mode of expression of the official temperature forecasts.

It might be argued that the uncertainty in temperature forecasts could be estimated from the mean and variance of an appropriate Gaussian (or normal) probability distribution, thereby eliminating the need for explicit assessment of probabilities for classes or intervals. However, the results presented in Section 5 indicate that 25% of the VW intervals were asymmetric. Moreover, the example of the guidance information provided to KNMI forecasters on the basis of the analogue procedure (see Table 1) reveals that the objective probabilistic temperature forecasts can be bimodal. Obviously, such forecasts cannot be modeled ade-

quately using a normal distribution. Thus, the assessment of probabilities for temperature classes or intervals is required, and the results of this study indicate that reliable and skillful probability forecasts of several different types can be prepared objectively. The choice of a particular format for such forecasts will depend on the relative utility of these formats, both to the forecasters and to the ultimate users.

Acknowledgments. The work reported here was initiated during the period June–September 1981 when the second author (AHM) was a visiting scientist at the Royal Netherlands Meteorological Institute. His participation in this study also was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grants ATM-8004680 and ATM-8209713. The authors would like to acknowledge the helpful comments of two referees.

REFERENCES

Balzer, K., 1976: Method and results of interpretation of geopotential forecasts. *Preprints, WMO Symp. on Interpretation of Broad-Scale NWP Products for Local Forecasting Purposes* (Warsaw). Geneva, WMO No. 450, 100–106.

Baur, F., 1951: Extended-range weather forecasting. *Compendium of Meteorology,* T. F. Malone, Ed., Amer. Meteor. Soc., 814–833.

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

Carter, G. M., J. P. Dallavalle, A. L. Forst and W. H. Klein, 1979: Improved automated surface temperature guidance. *Mon. Wea. Rev.,* **107,** 1263–1274.

Conte, M., C. De Simone and C. Finizio, 1980: Post-processing of numerical models: Forecasting the maximum temperature at Milano Linate. *Rev. Meteor. Aeronautica,* **40,** 247–265.

Daan, H., 1980: Climatology and persistence as reference forecasts in verification scores. *Preprints, Symp. on Probabilistic and Statistical Methods in Weather Forecasting* (Nice). Geneva, WMO, 195–201.

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

Duvernet, F., and D. Rousseau, 1980: Statistical interpretation of numerical weather prediction in the French Meteorological Service. *Preprints, Symp. on Probabilistic and Statistical Methods in Weather Forecasting* (Nice), Geneva, WMO, 397–400.

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

Francis, P.E., A. P. Day and G. P. Davis, 1982: Automated temperature forecasting, an application of Model Output Statistics to the Meteorological Office numerical weather prediction model. *Meteor. Mag.,* **111,** 73–87.

Glahn, H. R., 1983: Statistical weather forecasting. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, Boulder, CO (in press).

Klein, W. H., 1969: The computer's role in weather forecasting. *Weatherwise,* **22,** 195–218.

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

——, and R. L. Winkler, 1974: Credible interval temperature forecasting: Some experimental results. *Mon. Wea. Rev.,* **102,** 784–794.

——, and ——, 1979: Probabilistic temperature forecasts. The case for an operational program. *Bull. Amer. Meteor. Soc.,* **60,** 12–19.

Namias, J., 1951: General aspects of extended-range forecasting. *Compendium of Meteorology,* T. F. Malone, Ed., Amer. Meteor. Soc., 802–813.

Wilson, L. J., and N. Yacowar, 1980: Statistical weather element forecasting in the Canadian Weather Service. *Preprints, Symp. on Probabilistic and Statistical Methods in Weather Forecasting,* (Nice), Geneva, WMO, 401–406.

Winkler, R. L., 1972: A decision-theoretic approach to interval estimation. *J. Amer. Statist. Assoc.,* **67,** 187–191.

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

Woodcock, F., 1980: On the use of analogues to improve regression forecasts. *Mon. Wea. Rev.,* **108,** 292–297.

Yacowar, N., 1975: Probability forecasts using finely-tuned analogues. *Preprints, Fourth Conf. on Probability and Statistics in Atmospheric Sciences.* Tallahassee, Amer. Meteor. Soc., 49–50.