A Verification of Monthly Weather Forecasts in the Seventies

J. L. NAP, H. M. VAN DEN DOOL AND J. OERLEMANS

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

(Manuscript received 2 July 1980, in final form 23 October 1980)

ABSTRACT

Monthly forecasts of temperature, rainfall and sunshine have been verified during the period 1970–79. The predictions were based on seven different schemes. Of the seven methods, five refer to De Bilt (The Netherlands), one to southeast England and one to the Federal Republic of Germany. The results are not very encouraging for any of the methods. The skill is negligible except for a few schemes that predicted the monthly mean temperature ~ 10% better than climatology.

1. Introduction

At many places in the world long-range weather forecasts have been released over the years by national meteorological institutes. Most of the methods involved are a mixture of statistical and synoptic arguments. Although objective tools are used, the final forecast is subjective rather than objective. This means that a forecast made by A will not exactly be reproduced by an independent forecaster B, and in many cases it is even difficult to describe how the forecast is made.

It is well known that the skill of long-range forecasts is not very high. But even when the skill is said to be low, potential users are very eager to consult monthly or seasonal outlooks. This simple fact is the main reason that the release of forecasts goes on. But how useful are present-day long-range forecasts?

A complete answer to this question requires much more than what we are going to offer in this paper. We will give the results of a verification of forecasts of monthly mean quantities in an area of western Europe. The results tell us essentially about the skill. However, the utility of forecasts also depends on the user and it is possible that forecasts with a low skill can be used for cost/benefit effective planning and decision making. We restrict ourselves here to presenting the verification scores only.

How skillful are monthly forecasts? To answer this question we will examine forecasts made in the decade 1970–79. Forecasts of monthly mean quantities are usually made at the end of the previous month. We will verify five forecasts valid for De Bilt (The Netherlands), namely, 1) the National Meteorological Center (Washington, DC) forecast; 2) an analogue method; 3) statistical transition rules; 4) rules based on sea-surface temperature anomalies; and, last but not least, 5) pure persistence. In addition, we verify for (part of) their own area the forecasts produced by the meteorological services of 6) the United Kingdom and 7) the Federal Republic of Germany.

An objective way of measuring the skill of a forecast is to compare the number of hits obtained with a certain method over a long period with climatological chance. For that purpose a forecast has to be a clear statement about the interval in which the weather element in question is supposed to fall in the next month. When this turns out to be correct, we count a hit; partial hits do not exist. Depending on the width of the interval, randomly made forecasts will also score hits. The latter number of hits is proportional to the climatological chance C. After M forecasts with N hits the score is defined as

$$S = \frac{N}{M} - C. \qquad (1)$$

S measures the success expressed as percent better than chance. A value of 25% means that per 100 forecasts 25 more hits were obtained than can be expected on the basis of climatological chance.

For a large value of M the score S will distinguish between random forecasts ($S = 0$) and forecasts based on insight ($S > 0$). Of course, in a limited sample of M forecasts S suffers from sampling fluctuations and the statistical significance of a particular value of S must be specified. The chance of a hit being the climatological chance C, the percentage of expected hits, is C with a standard deviation of $M^{-1}[C(1 - C)M]^{1/2}$. This holds for an alterative distribution if C does not vary too much. For $C = 0.5$ and $M = 120$ (10 years), the standard deviation is 4.5%. Thus if the forecast is made at random, there is a chance of 5% that the score S exceeds 9% in absolute value.

0027-0644/81/020306-07$05.75

In the following sections we will give a brief description of the seven forecast methods that we have verified. Unfortunately, the description will not enable the reader to understand exactly how the forecasts are made in the various meteorological centers. This unsatisfactory situation is partly due to the subjective character of the methods. The forecaster makes a choice out of many guides and auxiliaries to arrive at the best possible forecast. As a result, the method cannot always be defined clearly. Another problem is that the methods change continuously. For example, extended runs of numerical weather prediction model up to 10 or even 30 days have been introduced during the 1970's as one of the tools to make a monthly weather forecasts. Also, a full documentation of the methods cannot always be found in the literature. So the descriptions in the present paper have to be in general terms.

These methods include seven schemes that predict monthly mean temperature; four of them also predict monthly precipitation and only one gives a sunshine forecast. Due to the very low skill in test periods, sunshine predictions are usually not even considered any more. For all methods and weather elements we discuss the score S for the 120 forecasts.

In nearly all schemes the temperature, precipitation, etc., are categorized into terciles, quintiles, etc. Most common is the tercile partition A (above), N (normal) and B (below). For all methods we specify what kind of partition is used.

2. Methods

In this section we will describe briefly the methods investigated. Some of them hardly deserve the name method (persistence) whereas others are very difficult to define because of the subjective input. Nevertheless, we will try to give descriptions. The first five methods yield a forecast valid for, among other places, De Bilt. The last two methods give forecasts for the United Kingdom and the Federal Republic of Germany.

a. National Meteorological Center (United States)

The method of the U.S. National Meteorological Center is basically the one described by Namias (1953), although certain adaptions have been made. On the basis of a 15-day mean map, a 5-day forecast and climatology, a 30-day mean 700 mb map centered at "today" is composed. The observed trends are projected into the next month. The anomalies in the forecasted 700 mb map are shifted somewhat, if necessary, in order to be consistent with the preferred positions of anomalies over the Northern Hemisphere. Preferred positions are known from general circulation statistics. Once the 700 mb map is prepared the anomalous heights and winds are translated into anomalous weather at the ground. In this stage a large input of statistical auxiliaries, like contingency tables for all states of the United States, is used to arrive at a forecast. The forecast for the European area is based almost completely on the 700 mb map. The final weather forecasts are presented in the form of a map for the Northern Hemisphere. From those maps we took the temperature class (A, N or B; with climatological chances 30, 40 and 30%, respectively) and the precipitation class (heavy or light; 50, 50%) for De Bilt.

b. Long-running analogues (Royal Netherlands Meteorological Institute)

Long-running analogues have been used in the Dutch Weather Service as a method to predict monthly mean temperature, precipitation and sunshine. With the help of historical daily weather maps, years are selected in which the evolution of atmospheric circulation types over Europe had a certain degree of resemblance with the present year. Only when this resemblance lasts for at least a few months is the historical year qualified as being a long-running analogue. The available record spans the period 1881 to the present.

The main problem with this method is that there are never really good analogues. One then must deal with four or five mediocre analogues, which often show divergent behavior in their further development. The process of deriving a forecast from the analogues is highly subjective. The method of long-running analogs has been used to forecast temperature, the number of dry days and monthly sunshine, all in terms of a tercile partition. The method has been described in more detail by Schuurmans (1973).

c. Transition rules

Transition rules express the fact that the chances for a warm, cold or normal next month depend on the temperature of the past month. In many cases there will be tendency for a temperature anomaly to persist to the next month (Van den Dool and Nap, 1981). Such relations of temperature in adjacent months can be determined from the observed records at De Bilt. Fig. 1 shows as an example the transition from January to February. The tercile partition for February is indicated by horizontal lines. After a cold January, say, −1°C, the chances for February to fall in the A, N or B class are 25, 25 and 50%. After a warm January, say, 5°C, the chances are 40, 45, and 15% for A, N or B to occur.

The transition rules were derived for the period 1881–1969 and applied to the years 1970–79. The preferred classes (chance > 34%) were taken as the forecast. The same method has been described by Gordon and Wells (1976); they call it the optimum

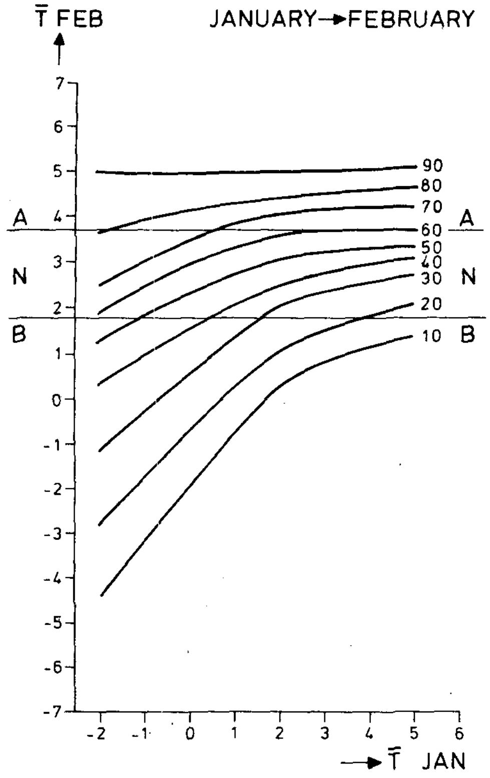

FIG. 1. Conditional cumulative probabilities of the mean temperature in February at De Bilt. In dependence on the temperature in January (horizontal axis) the contours measure the chance (%) that the temperature in February will not exceed the value given on the vertical axis. The tercile distribution used for the February temperature is indicated with horizontal lines. This diagram is based on observations during 1881–1969.

probable change method and report an application to the Central England series.

We will not discuss here transition rules for precipitation and sunshine, because even in the dependent data set they have hardly any skill.

d. Sea surface temperature rules

After the pioneering work of Namias (1965) and others, sea surface temperature anomalies (SSTA) are used in long-range weather forecasting. The monthly mean circulation over Europe has been shown to correlate with SSTA in the North Atlantic, both in terms of pressure patterns (Ratcliffe and Murray, 1970) and circulation types (Oerlemans, 1975).

The current method is based on a part of the SSTA classification of Ratcliffe (1971) and on the temperature record of De Bilt. Based on the 1877–1969 period, conditional probabilities have been computed for the monthly temperature T of the next month to be in the A, N or B class. In the case of a cold pool (CP) near New Foundland in May, the chance of T to be in the A, N or B class in June is 45, 40 and 15%, respectively. The chances differ from month to month, even if the SSTA is the same. The SSTA are divided into three categories: CP, WP (warm pool near New Foundland) and neither CP nor WP (cold pool east and cold pool west fall in the CP class, warm pool east and warm pool west in the WP class). The forecast for T consists of stating that T will be in the class(es) with $P > 34\%$.

e. Persistence

Persistence in a pure form can be used as a reference for skill, it is a gift of nature. Thus one applies $A_i \rightarrow A_{i+1}$, $N_i \rightarrow N_{i+1}$ and $B_i \rightarrow B_{i+1}$, where the index i is number of the month. This appears to be a good forecast in some periods of the year and a mediocre one in other periods (Van den Dool and Nap, 1981).

f. United Kingdom

Like the NMC method, the United Kingdom monthly forecast is also obtained by adding and integrating various indications for the next month's circulation. Those indications come from historical analogues, sea surface temperature, extrapolation of circulation index series, etc.

The forecast is made for different parts of the United Kingdom. We only verified the forecasts for southeast England. The temperature forecast indicates one or more preferred classes out of five classes, which all have a climatological chance of 20%. For precipitation three classes are used.

g. Federal Republic of Germany

The monthly forecast issued by the German Weather Service at Offenbach is based on a multiple-regression technique, with a large number of predictors. Historical analogues also play a role. The forecasts are formulated in terms of temperature and precipitation intervals, and the classes are not fixed. The verification presented here gives essentially a skill score averaged over the country.

3. Results

a. Overall skill

Fig. 2 displays the score of the various methods in forecasting monthly mean temperature. Both yearly and cumulative values of S are given.

It appears that the overall score is positive, but that only two methods are better than "chance" at the 95% confidence level, *viz.*, transition rules and sea surface temperature rules. They have average scores of 9 and 11%, respectively, and seem to be better than persistence which has a score of 5% over climatology.

Fig. 3 shows in the same format the scores for precipitation and sunshine. Only one comment can be

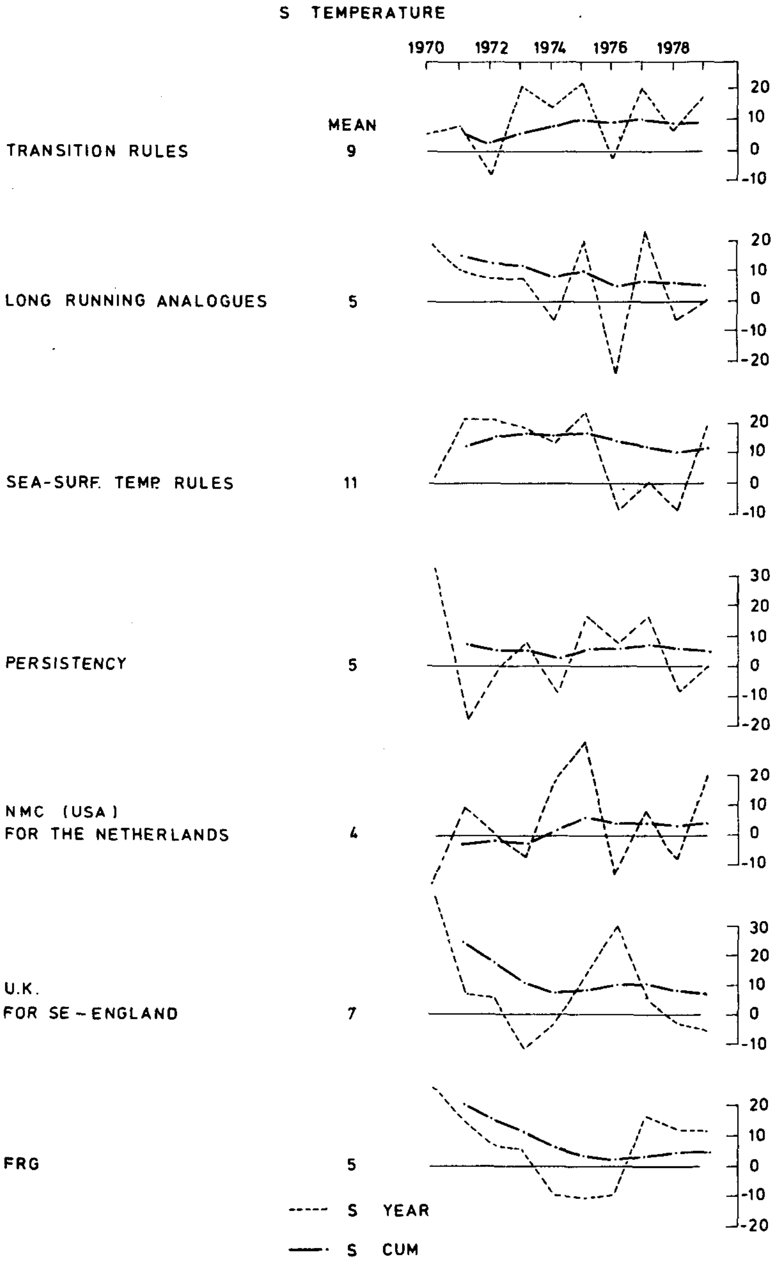

FIG. 2. The verification score S measuring the skill of monthly mean temperature forecasts. From top to bottom the results refer to methods that yield a forecast for De Bilt, while the lowest two are for southeast England and the Federal Republic of Germany, respectively. In each block the dashed lines connect values of S per year (12 forecasts) while the dashed-dotted line represents the cumulative value of S.

given to this figure: monthly forecasts of precipitation and sunshine did not have any skill.

b. *Skill for months with extreme temperature or precipitation*

The methods discussed here do not really forecast extremes but it is interesting to know whether the score averaged over extreme months is higher or lower than the total score. To investigate this we define an extreme month as a month in which the mean temperature (or precipitation) deviates more than 1.5 times the standard deviation from the mean. Applying this requirement to De Bilt, it appears that there were 11 extreme temperature months and 18 extreme precipitation months between 1970 and 1979. The scores are listed in Table 1.

Although sampling errors may be large now, the

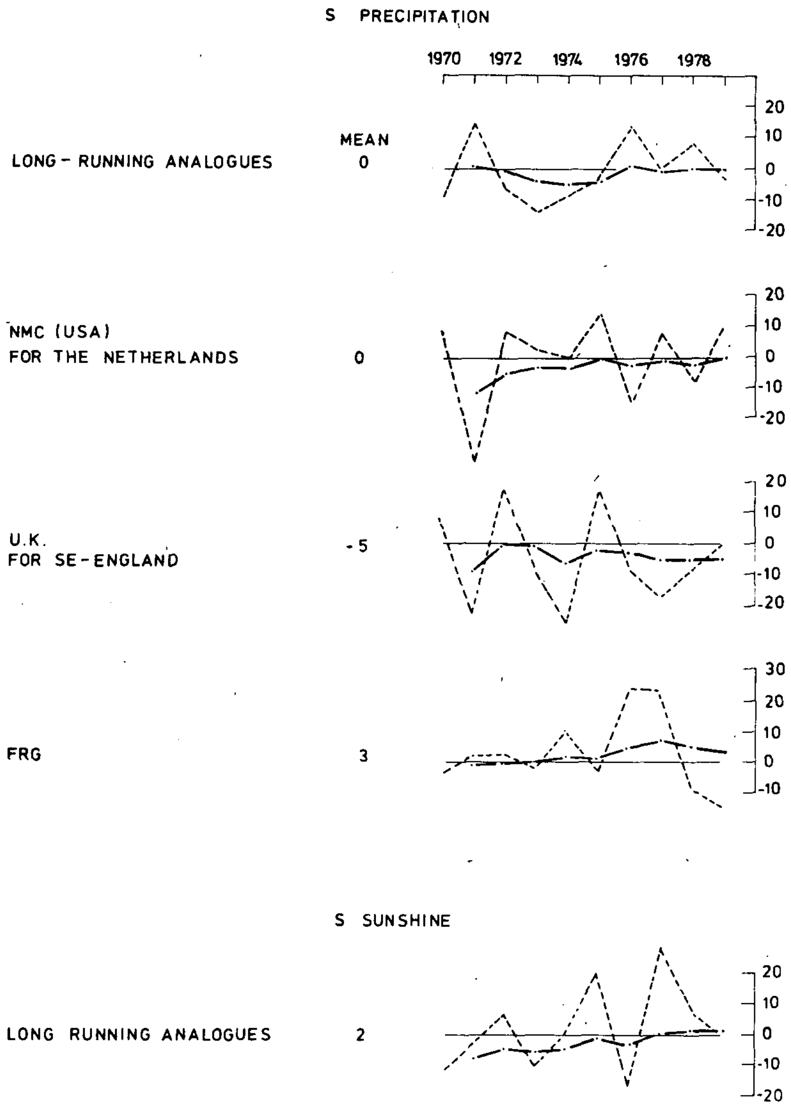

FIG. 3. As in Fig. 2 but for precipitation and sunshine.

figures strongly suggest that persistence performs best. Apparently, a very warm month tends to follow a warm one rather than a normal or cold one. One should be careful in interpreting this result, however. It does not mean that persistence predicts extremes very accurately, only that the (broad) class is predicted better than by other methods.

c. *Seasonal variation of S*

Another question is whether S shows a large variability through the year or not. For one specific method, the sample is too small to investigate this. We therefore put together all temperature forecasts and arrived at the distribution shown in Fig. 4.

Late summer and late winter appear to be the best predictable periods of the year. In general, these are the periods in which persistence of temperature in northwestern Europe is highest (Van den Dool and Nap, 1981).

d. *Mutual dependence of transition and sea surface temperature rules*

Transition rules (TRU) and sea surface temperature rules (STRU) both seem to yield a positive skill. It is important to know whether those methods contain essentially the same information or not. Or, in other words, do they obtain their hits in the same months? Let us consider this more closely.

If for TRU and STRU the chance of a hit is p and q, respectively, the chance of two hits or two failures is $P = 2pq + 1 - p - q$. In the present case we have

$p = 0.58$ and $q = 0.62$, which yields $P = 0.52$. We therefore expect that in 62.4 out of the 120 cases TRU or STRU are both successful or both fail. In reality, this was the case in 68 months, which means that no strong dependence exists between the methods. By using an optimum combination, it should thus be possible to increase S by a few percent.

Because scores are so small, we did not spend much effort in analyzing the properties of S for other methods. As Fig. 2 shows, there seem to be good and bad years, but the overall picture is noisy.

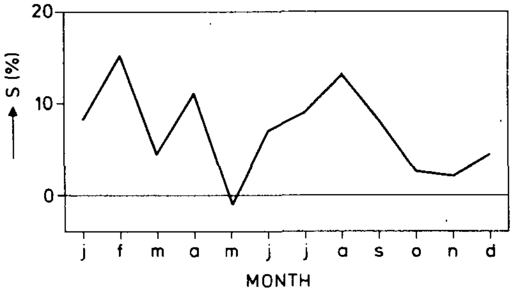

FIG. 4. The skill of monthly mean temperature forecasts averaged over all methods. The value of S (%) is given as a function of the target month of the forecast.

4. Discussion

The picture emerging here is rather disappointing: we have to conclude that, at least for Western Europe, monthly forecasts were not very successful during the 1970's. Only for mean temperature was some skill present.

Another conclusion to be drawn is that simple statistical methods (transition rules), which require a 1 min effort to make a forecast, perform at least as well as more sophisticated and time-consuming methods (analogues).

These conclusions are not entirely new. Both the producers and users know that monthly forecasts have a low skill. Gordon and Wells (1976) used an optimum probable change method which, applied to the central England temperature series, yields a score $S = 6\%$. In the same paper the score of the official monthly temperature forecast for the whole United Kingdom is reported to be ~5% over a 12-year period. These verification scores are consistent with ours.

Although the skill of the NMC temperature forecast is not very high for De Bilt, the score S (4%) is significantly larger than zero when we extend the period of verification to 1954–79. For the United States the forecasts based on the same 700 mb maps turn out to be better (Gilman, personal communication, 1978). This is due to the additional input during the process of translating the 700 mb map to weather in the United States.

During the 1970's several advances were made in meteorology that may ultimately improve long-range weather forecasting. The first is the slow but continuous improvement of the numerical weather prediction models. For the first ten days of the coming month such models may give the best possible guide and also the averaged model output from 10 to 30 days may become of great value for the prediction of statistics of the weather in that period. A second step forward is the observed connections between SST and the future atmosphere. In fact, the sea surface temperature rules, described in Section 2, are based on connections proposed by Ratcliffe and Murray (1970). New evidence of SSTA-atmosphere relations is claimed by many workers and applied with some success by Harnack (1979) in the statistical prediction of winter temperatures in the United States. Although these developments are promising it is far from certain that it will raise the level of skill of long-range weather forecasts very much. Looking at the results presented in this paper we cannot avoid the conclusion that there has been little progress in making monthly weather forecasts for northwestern Europe. This supports the idea of some workers that significant progress is impossible. Long-range forecasts are possible only if, in addition to the noise of unpredictable weather systems, there are deviations from normal caused by well-defined anomalies in the boundary conditions (the signal). Estimates of the signal-to-noise ratio by Madden (1976) support the pessimistic view that the signal is generally very weak at midlatitudes. Therefore, it is quite possible that verification scores on the order of 10–20% are a fundamental maximum rather than the present state of the art. Even though this may seem a low level of skill, it may be that for certain purposes long-range weather forecasts are useful.

TABLE 1. Skill score of various methods for extreme months (%).

	Temperature	Precipitation
Transition rules	21	
Sea surface temperature rules	24	
Long-running analog	−36	−26
National Meteorological Center	14	4
Persistence	49	15

Acknowledgments. The verification of one single forecast is a 1 min job. The value of this paper is determined by the length and enduring character of the verification experiment. In earlier years the experiment had been conducted by Cor Schuurmans, Henk Krijnen, Hans Reiff and Fons Baede. We thank Mr. D. Houghton (UK), Mr. H. Reiser (FRG) and the reviewers for their critical reading of the manuscript.

REFERENCES

Gordon, A. H., and N. C. Wells, 1976: Changes in temperature from month to month for central England for a quintile distribution. *J. Appl. Meteor.*, **15,** 928–932.

Harnack, R. P., 1979: A further assessment of winter temperature predictions using objective methods. *Mon. Wea. Rev.*, **107,** 250–267.

Madden, R. A., 1976: Estimates of the natural variability of the time-averaged sea-level pressure. *Mon. Wea. Rev.*, **106,** 279–295.

Namias, J., 1953: *Thirty-Day Forecasting: A Review of a Ten-Year Experiment. Meteor. Monogr.*, No. 6, Amer. Meteor. Soc., 83 pp.

——, 1965: On the nature and the cause of climatic fluctuations lasting from a month to a few years. WMO Tech. Note No. 66, 46–62.

Oerlemans, J., 1975: On the occurrence of "Grosswetterlagen" in winter related to anomalies in North Atlantic sea temperature. *Meteor. Rdsch.*, **28,** 83–88.

Ratcliffe, R. A. S., 1971: North Atlantic sea temperature classification. *Meteor. Mag.*, **100,** 225–232.

——, and R. Murray, 1970: New lag associations between North Atlantic sea temperature and European pressure applied to long-range weather forecasting. *Quart. J. Roy. Meteor. Soc.*, **96,** 226–246.

Schuurmans, C. J. E., 1973: A 4-year experiment in long-range weather forecasting using circulation analogues. *Meteor. Rdsch.*, **26,** 2–4.

Van den Dool, H. M., and J. L. Nap, 1981: An explanation of persistence in monthly mean temperatures in The Netherlands. *Tellus*, **33** (in press).