ETDL 79-001

02362
02353

U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
SYSTEMS DEVELOPMENT OFFICE
TECHNIQUES DEVELOPMENT LABORATORY

TDL OFFICE NOTE 79-1

COMPARATIVE VERIFICATION OF OPERATIONAL TWO TO SIX HOUR OBJECTIVE FORECASTS AND OFFICIAL NWS WATCHES OF SEVERE LOCAL STORMS: AN UPDATE

Jerome P. Charba and Stephen M. Burnham

January 1979

COMPARATIVE VERIFICATION OF OPERATIONAL TWO TO SIX HOUR OBJECTIVE FORECASTS AND OFFICIAL NWS WATCHES OF SEVERE LOCAL STORMS: AN UPDATE

Jerome P. Charba and Stephen M. Burnham

1. INTRODUCTION

In a previous article (Charba and Burnham, 1978), we described a procedure used to perform a comparative verification of 2-6 h objective forecasts of severe local storms developed by the Techniques Development Laboratory (TDL) and tornado and severe thunderstorm watches issued by the National Severe Storms Forecast Center (NSSFC). The article included verification scores for the operational forecasts issued by each system during the spring season of 1977. The present note essentially updates the verification scores of 1977 with those from the spring season of 1978. It also considers a factor which could explain the improvement found in the scores of both systems from 1977 to 1978.

2. BACKGROUND

The procedure used to perform the comparative verification is described in detail in Charba and Burnham (1978) (henceforth referred to as A). While the reader should refer to A for details, a brief summary of the procedure is included here to provide a framework for discussion of the results presented.

The forecasts included in the verification consisted of the operational 2-6 h objective forecasts and all tornado and severe thunderstorm watches which fell within the interior area delineated by the dotted line in Fig. 1. Any forecast whose valid period fell totally within or had a minimum overlap of one hour with the period 2000-0600 GMT was included in the verification. All daily forecasts produced by each system which fell within this temporal and spatial domain for the period 16 March to 15 June 1978 were verified together.

The scores computed from the sample of forecasts for each system are the probability of detection (POD) and the false alarm ratio (FAR) (see Donaldson et al., 1975). In order to compute these scores, the objective forecasts, which were issued in probability form, are converted into categorical YES/NO forecasts and verified over small square areas. Watch rectangles were broken down into square areas of the same size and each such square was defined to be a separate YES forecast (Fig. 1). The POD and FAR were then computed from standard 2x2 contingency tables.

While the POD and FAR properly portray each system's ability to forecast occurrences of severe storms, these scores take no account of the lead times and valid periods of the forecasts. In A, we showed that there were significant differences in lead times and valid periods between the two sets of forecasts. Augmented verification results which illustrate these differences are shown in Figs. 2a and 2b. [These figures are identical to those contained in A except that the data from the new season (1978) were added to those used previously (1977) to form the sample for each system.] Fig. 2a shows frequency histograms of storm occurrences correctly forecast (hits) relative to all storm occurrences

as a function of time after forecast issuance. This figure also shows the average lead time, valid period, and projection time (i.e., storm occurrence time relative to the beginning of the valid period) corresponding to these hits. Fig. 2b shows similar results for correct YES forecasts. The frequencies plotted in Fig. 2b are different from those in Fig. 2a, however, in that the number of correct YES forecasts in a half-hour interval is divided by the number of YES forecasts valid in the interval; thus, the frequency may be considered as the percent correct. Division by the number of valid YES forecasts is necessary because this number may change from one half hour to the next in the case of the watches.

The main point to note from Figs. 2a and 2b is that the lead times and valid periods for hits and correct YES forecasts are significantly different between the two forecasting systems. The lead time is longer and the valid period is shorter for the objective forecasts as compared to averages of these parameters for the watches. It is also interesting to note from the watch frequency histograms that while the number of hits in the first half-hour interval after issue time is relatively small (Fig. 2a) the number of valid YES forecasts is also small; thus, the percent correct is conspicuously high. This indicates that it is easier for NSSFC forecasters to correctly forecast storms which occur soon after forecast issuance (and, thus, with little or no lead time) as compared to those which occur several hours later. While the tendency for relatively high percent correct early in the valid period is also apparent for the objective forecasts (Fig. 2b), a lead time of 1 1/4 hours exists for all of these forecasts. All of these results further illustrate the need for taking the lead time and valid period into account in the scoring as was previously recognized in A.

In A, a scheme was developed to account for these time parameters in the computation of the POD and FAR in a reasonable manner. Following this scheme, each correct forecast is given a weight, W, defined as:

$$W = \alpha \frac{LT}{VP} + \beta \frac{PT}{VP} ,$$

where LT, VP, and PT are the lead time, valid period, and projection time. The empirical factors α and β were assigned the values 2.0 and 1.0, thus giving greater importance to lead time over projection time. The weight function W is incorporated in the definition of the "weighted" POD and FAR as follows: each hit or correct YES forecast is assigned the value of its weight (W) instead of the value of one which is effectively given when the standard POD or FAR is computed from 2x2 contingency tables.

In the next section, we present verification results from the spring season of 1978 using both the weighted and unweighted POD and FAR. Corresponding results previously presented for the spring season of 1977 are also included here for the purpose of comparison.

3. RESULTS

First, we shall examine how each system performed in correctly forecasting the individual storm occurrences that fell within the temporal and spatial domain involved in the verification. Table 1a contains the number of storm

occurrences, the number of hits, and the unweighted POD scored by each set of forecasts in the Gulf and non-Gulf regions (Fig. 1). Corresponding results for the 1977 season are given in parentheses. Note that the POD values for the watches are much larger than are those for the objective forecasts. The superior performance of the watches is especially prominent in the Gulf region. It is also noteworthy that the scores from each system improved considerably from 1977 to 1978, with slightly greater improvement being shown by the objective system.

Table 1b gives the number of YES forecasts, the number correct, and the FAR for each of the two systems. Note that the relative skill of the two systems expressed in terms of the FAR is similar to that shown in the case of the POD. That is, the watches produced a better FAR in each region, with the margin of superiority being greater in the Gulf region. Also note the improvement in FAR for both systems from 1977 to 1978.

Tables 2a and 2b contain information similar to Tables 1a and 1b, the essential difference being that the individual hits and correct YES forecasts were weighted according to the scheme described earlier. Also given is the average weight over all hits and correct YES forecasts. Table 2a shows that the weighted POD's for the combined regions were the same for the two systems--a remarkable coincidence! As for the individual regions, the objective system had a higher weighted POD in the non-Gulf region while the watches retained their advantage in the Gulf region. Table 2b shows that in 1978 the weighted FAR for the objective system was slightly better than that for the watches in both the Gulf and non-Gulf regions. This result differs slightly from that seen for 1977 in that, for this previous season, the watches retained a slight edge in the Gulf region. The reason for the improved performance of the objective forecasts relative to the watches seen in these weighted scores can be noted from the right hand columns of Tables 2a and 2b. The average weight values corresponding to the objective forecasts were almost twice as large as those for the watches. This result reflects the large differences in lead time and valid period between the two systems as discussed earlier.

4. DISCUSSION

One item that stands out in Tables 1a, 1b, 2a, and 2b is that the forecast performance of each system improved considerably from 1977 to 1978. While an analysis of all factors which may have contributed to this improvement is beyond the scope of this study, we did attempt to determine if the objective system used in 1978 was more skillful than that used in 1977. In particular, since interactive predictors (see National Weather Service, 1978) were introduced to the system immediately prior to the 1978 season, we wanted to determine how much (if any) of the improvement seen in Tables 1 and 2 was due to this change. We proceeded by applying the operational system used in 1977 to an archive of the real time data of 1978. The resulting sample of forecasts was then verified in the same manner as the operational forecasts of 1978, i.e., as described in this note. Care was taken to insure that the data sample extracted from the archive was identical to that which was used for the operational forecasts.

To our surprise, the results of this test showed that the scores achieved by the 1977 system on the 1978 data were almost identical to those produced by the operational system of 1978. Therefore, in terms of the categorical forecasts used in this comparative verification, the improved scores exhibited by the operational objective forecasts in 1978 could not be attributed to an actual improvement in the forecasting system over that used in 1977. Thus, we are left with the conclusion that forecasting for the 1978 season was easier for the objective system.

Since the objective forecasts are operationally available to NSSFC forecasters as guidance and since the scores of both systems improved from 1977 to 1978, one has to wonder how much (if any) impact the improvement in objective forecasts had on the improvement of the watches. In this regard, it is interesting that the relative skill of the two systems was about the same in 1977 and 1978. Whether this parallel improvement is due to substantial use of this guidance product by NSSFC forecasters or whether it only reflects the difference in predictability during the two seasons, with the objective forecasts having no impact on the watches, we can't say. The true situation likely falls somewhere between these two extremes.

Finally, it is important to bear in mind that these forecasts are operationally available to NSSFC in probability form. Therefore, it is more relevant to consider the relative skill of the 1977 and 1978 objective systems in terms of the P-score as computed from the probabilities. We computed the P-score (actually one-half the P-score defined by Brier, 1950) achieved by the 1977 and 1978 objective systems for the spring season data sample of 1978. The results showed that the P-score of the 1978 system was 1.5% better (lower) than that of the 1977 system. Thus, while the categorical forecasts from the 1978 system were not better than those from the 1977 system, the corresponding probability forecasts produced by this system did perform slightly better than those of its predecessor. Of course, we would like to think that this increase in skill of the objective system contributed in some way toward the improvement in the watches in 1978.

5. SUMMARY

The enlarged sample involved in this comparative verification of TDL's 2-6 h objective guidance forecasts and official watches of severe storms issued by NSSFC has produced the following results:

1) The watches scored substantially better than the objective forecasts when the lead times and valid periods of both sets of forecasts were not taken into account in the scoring.

2) The watches had substantially shorter lead times and longer valid periods than did the objective forecasts. Thus, there was a clear need for incorporating these time parameters in the development of additional scores.

3) Scores which took account of the lead times and valid periods in a reasonable manner showed that the watches and objective forecasts performed at about an equal level of skill.

4) There was substantial improvement in the scores of both the objective forecasts and the watches from the spring of 1977 to the spring of 1978.

5) The objective system used in 1978 was no better than the objective system used in 1977 in terms of its ability to produce correct categorical forecasts. However, the 1978 objective system was slightly more skillful than its predecessor in terms of the probability forecasts, the form in which these forecasts are available to field forecasters.

5. ACKNOWLEDGEMENTS

We thank Allen Pearson of NSSFC for providing the severe storm and watch data. We also thank the following TDL members for their contributions: William Griner for helping develop some of the computer programs, Denis Sakelaris for assisting with the illustrations, and Peggy Gardner for typing the manuscript.

6. REFERENCES

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

Charba, J. P., and S. M. Burnham, 1978: Comparative verification of operational two to six hour objective forecasts and official NWS watches of severe local storms. Conference on Weather Forecasting and Analysis and Aviation Meteor., October 1978, Silver Spring, Md., Amer. Meteor. Soc., 20-27.

Donaldson, R. J., Jr., R. M. Dyer, and M. J. Kraus, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints 9th Conference on Severe Local Storms, October 1975, Norman, Okla., Amer. Meteor. Soc., 321-326.

National Weather Service, 1978: Two to six hour probabilities of thunderstorms and severe local storms. NWS Technical Procedures Bulletin No. 228, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 13 pp.

Table 1a. Number of hits and probability of detection (POD) scored by the objective forecasts and watches over all storm occurrences falling within the verification domain for the spring season (16 March to 15 June) of 1978. Results for the 1977 season are given in parentheses. See text for further details.

FORECASTS	REGION	OCCURRENCES	HITS	POD
OBJECTIVE	GULF	225 (135)	50 (13)	0.22 (0.10)
OBJECTIVE	NON-GULF	648 (701)	246 (179)	0.38 (0.26)
OBJECTIVE	COMBINED	873 (836)	296 (192)	0.34 (0.23)
WATCHES	GULF	225 (135)	125 (44)	0.55 (0.33)
WATCHES	NON-GULF	648 (701)	375 (297)	0.58 (0.42)
WATCHES	COMBINED	873 (836)	500 (341)	0.57 (0.41)

Table 1b. The number of correct severe storm (YES) forecasts and FAR scored by each system for all YES forecasts falling within the verification domain for the spring season of 1978. Results for the 1977 season are given in parentheses.

FORECASTS	REGION	YES FORECASTS	CORRECT	FAR
OBJECTIVE	GULF	288 (186)	29 (8)	0.90 (0.96)
OBJECTIVE	NON-GULF	1277 (1484)	165 (135)	0.87 (0.91)
OBJECTIVE	COMBINED	1565 (1670)	194 (143)	0.88 (0.91)
WATCHES	GULF	392 (369)	68 (39)	0.83 (0.89)
WATCHES	NON-GULF	1556 (1511)	242 (202)	0.84 (0.87)
WATCHES	COMBINED	1948 (1880)	310 (241)	0.84 (0.87)

Table 2a. Same as Table 1a except that the hits and the POD are weighted according to the lead times and valid periods of the forecasts as described in the text. The average weight over all hits is given in the right hand column.

FORECASTS	REGION	OCCURRENCES	HITS	POD	AVG. WEIGHT
OBJECTIVE	GULF	225 (135)	54.7 (15.3)	0.24 (0.11)	1.10 (1.18)
OBJECTIVE	NON-GULF	648 (701)	275.0 (196.3)	0.43 (0.28)	1.12 (1.10)
OBJECTIVE	COMBINED	873 (836)	329.7 (211.6)	0.38 (0.25)	1.12 (1.11)
WATCHES	GULF	225 (135)	72.7 (27.3)	0.32 (0.20)	0.58 (0.62)
WATCHES	NON-GULF	648 (701)	259.6 (181.2)	0.40 (0.26)	0.69 (0.61)
WATCHES	COMBINED	873 (836)	332.3 (208.5)	0.38 (0.25)	0.67 (0.61)

Table 2b. Same as Table 1b except that the correct YES forecasts and FAR are weighted according to the lead times and valid periods of the forecasts (see text). The average weight over all correct YES forecasts is given in the right hand column.

FORECASTS	REGION	YES FORECASTS	CORRECT	FAR	AVG. WEIGHT
OBJECTIVE	GULF	288 (186)	29.6 (8.0)	0.89 (0.96)	1.02 (1.00)
OBJECTIVE	NON-GULF	1277 (1484)	169.6 (139.6)	0.87 (0.91)	1.03 (1.03)
OBJECTIVE	COMBINED	1565 (1670)	199.2 (147.6)	0.87 (0.91)	1.03 (1.03)
WATCHES	GULF	392 (369)	36.2 (24.4)	0.91 (0.93)	0.53 (0.63)
WATCHES	NON-GULF	1556 (1511)	152.7 (111.8)	0.90 (0.93)	0.63 (0.55)
WATCHES	COMBINED	1948 (1880)	191.9 (136.2)	0.90 (0.93)	0.61 (0.56)

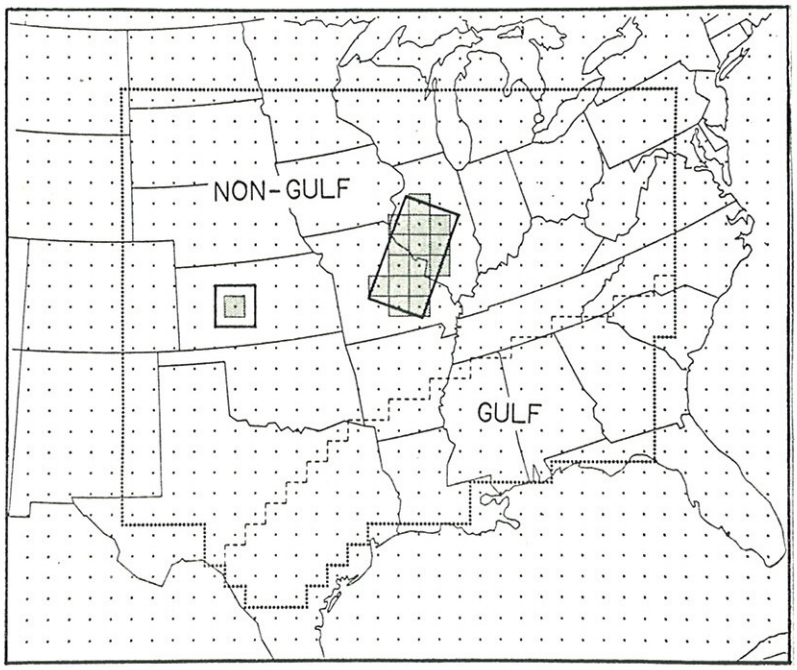

Figure 1. Areas involved in the comparative verification. The total area, enclosed by the heavy dotted line, is divided into "Gulf" and "non-Gulf" regions as delineated by the irregular dashed line. The area over which an individual objective (probability) forecast is valid is illustrated by the larger square area in Kansas. The corresponding categorical YES/NO forecast is verified over the shaded square centered inside the larger one. The example watch area is broken into square areas of the same size and each is verified as a separate YES forecast.

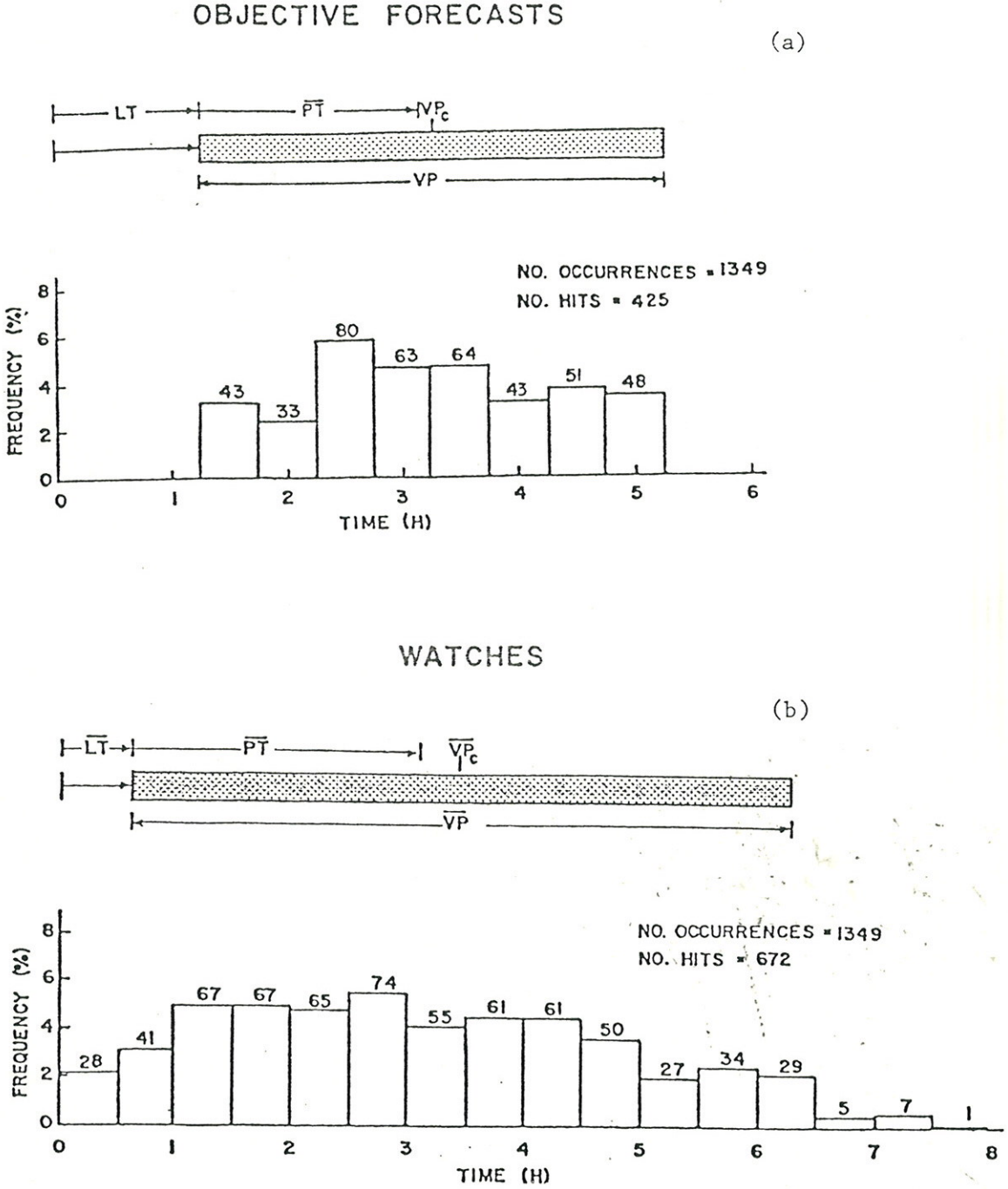

Figure 2. Frequency of storm occurrences correctly forecast (hits) in the non-Gulf region relative to all storm occurrences in that region as a function of time after forecast issuance. Fig. 2a pertains to the objective forecasts and Fig. 2b is for the watches. The actual number of hits within each half-hour interval is shown. The schematic above each histogram depicts the average lead time ($\overline{LT}$), average valid period ($\overline{VP}$), and average projection time ($\overline{PT}$) over all hits. $\overline{VP}_c$ denotes the center of the $\overline{VP}$. Since the LT and VP for the objective forecasts are fixed, bars do not appear with these symbols in (a).

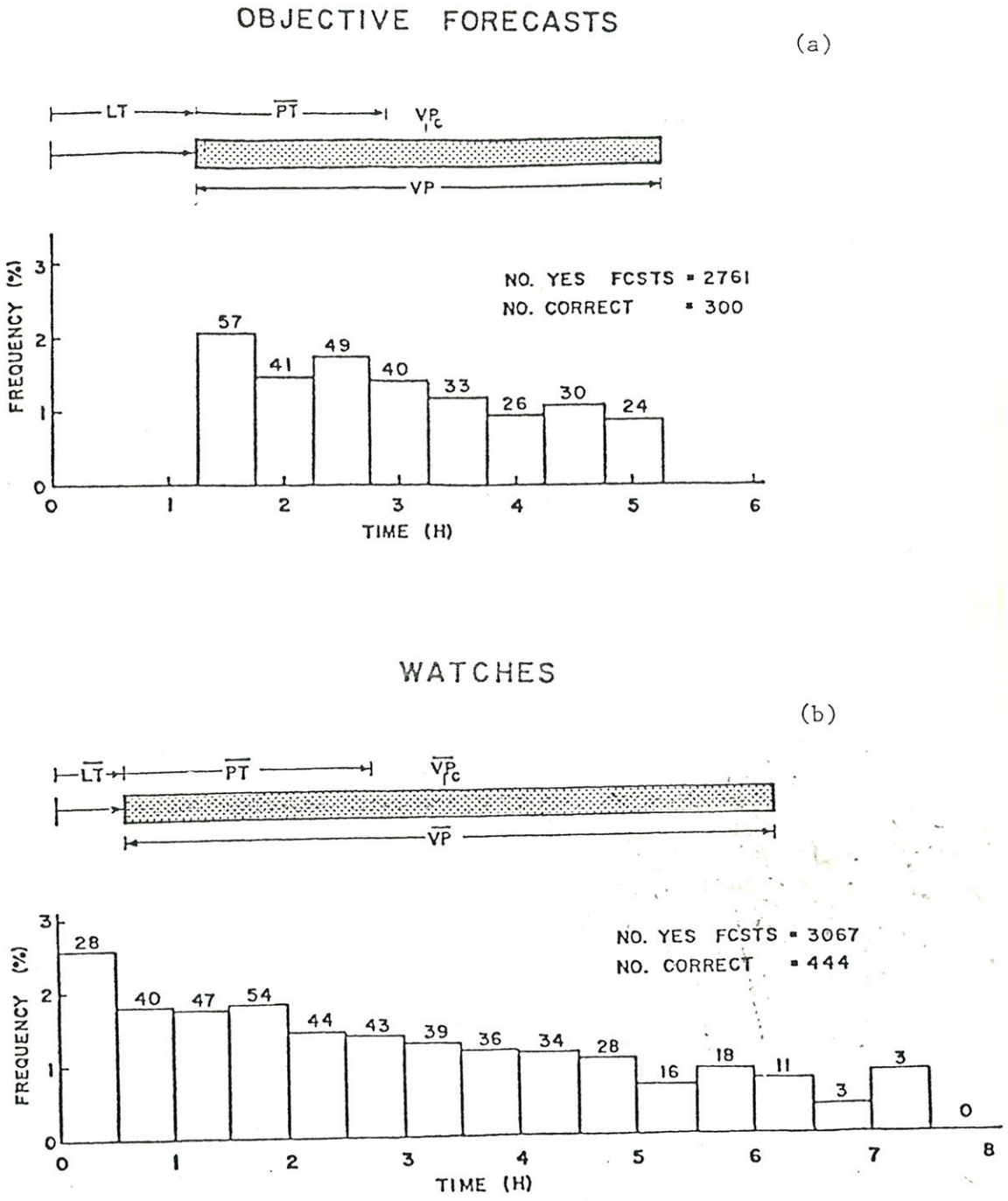

Figure 3. Similar to Fig. 2 except for YES forecasts. The frequency for each half-hour interval is the number of correct YES forecasts divided by the number of YES forecasts valid in the interval. Note that in the case of the watches (Fig. 3b) the number of valid YES forecasts may vary from one half hour to the next.