PACIFIC REGION TECHNICAL NOTES

79-017

June 6, 1979

Verification Experiments at Comox

W.L. Ranahan

Introduction

Verification of terminal or other forecasts has been a difficult task due to the limitations of the various verification methods to date, and the problems in interpretation once results are obtained. The discussion by Reid (1) in a recent edition of Atmosphere-Ocean encouraged some meteorologists to experiment with the Ranked Probability Score (RPS). Weather Centres equipped with computers could do this on a routine basis with near-real-time feedback. Another paper presented at the Conference on Weather Forecasting and Analysis and Aviation Meteorology, Silver Spring, Md., in October 1978, dealt with the verification of weather watches and severe storms in the U.S. This method by Charba and Burnham (2) was adapted at Comox for a single station forecast of significant weather events and the verification of same. Results of the two experiments are discussed.

Ranked Probability Score (RPS)

The methods and weather categories used were identical to those used by Reid in assessing forecasts produced for Winnipeg at the Prairie Weather Centre. The period involved was from January to April 1979, and the first twelve hours of the official terminal forecast prepared by Comox meteorologists was verified. As with Reid's method, the comparative Persistance Forecast was simply an extension of the weather category observed on the hour prior to the twelve hour forecast. The climatology forecast was derived from the frequencies of occurrence of the six weather categories in each month over the last five years.

The results of the RPS experiment for the four month period were similar to those found by Reid at Winnipeg. Figure 1 is a graph showing the accuracy of the Comox terminal forecast, climatology forecast and persistance forecast as a function of time into the period of the forecast. From the graph, it can be seen that the performance of the terminal forecast marginally exceeded that of the persistance forecast even over the first hour or so. Reid found that the persistance forecast outperformed the terminal forecast for the first hour. Secondly, the terminal forecast accuracy exceeded that of the climatology forecast until after seven hours into the forecast period. At Winnipeg, the climatology forecast excelled over the terminal forecast after five hours. The individual scores among forecasting staff fell into a narrow range, from .945 to .956.

The better scores at Comox can probably be attributed to the single station terminal forecast responsibility, rather than the many terminals that the Prairie Weather Centre is responsible for. Although the Comox forecaster is preoccupied with many duties, he is able to concentrate on the accuracy of the terminal forecast for one station - his own. Another reason for the higher Comox scores may be the better ceilings and visibilities found at Comox on the average. Even RPS has some fine weather bias, although it was not measured.

The climatology score is somewhat inflated because it is not limited to specific values as are the terminal and persistance forecasts. A 60/40 judgement by the forecaster often ends up as a 50/50 (VRBL) forecast, where a 60/40 climatology statistic can be entered directly. The 1979 climatology forecast was adjusted as follows to consider the same limitations of the terminal forecast:

It would have been expected that the climatology forecast with the probability limitations would have scored poorer in the period than the direct entry of the five-year average frequencies. However, in 1979, this was not the case. This was due largely to the similiarity in frequencies between the statistics with the limitations, and the 1979 figures. In other years, the climatology RPS score would likely be lower.

In summary, the RPS method of verifying terminal forecasts is a valid verification tool. Nevertheless, it suffers from some of the same weaknesses as its predecessors. There would still be a tendency to bias in favor of the fine weather. In addition, there would be an encouragement to hedge for higher scores. The use of the 50/50 VRBL term in the case of forecasting COXOF VRBL/SCT would always result in a score of .75, too high for such a useless forecast. In this sense, the RPS method of verification could result in less useful forecasts.

POD and FAR

Probability of Detection (POD) and False Alarm Ratio (FAR) are an adaptation of severe weather forecast verification methods in the U.S. This experiment was very useful in evaluating forecast skill and usefulness. Significant, rather than severe, weather events were established to concern wind, aviation weather, rain and snow. Near the end of each shift, each forecaster would fill out a simple check-sheet on any significant weather events for the next 24 hours. At Comox, the following events were established as significant, reproduced on the checklist below:

FCST	ACTUAL	EVENT
		SFC WND 25 KT OR MORE
		SFC WND 40 KT OR MORE
		SFC WND 55 KT OR MORE
		6 HRS OR MORE VFR
		3 HRS OR MORE IFR
		10 MM OR MORE RAIN
		2-10 CM SNOW
		11-25 CM SNOW
		MORE THAN 25 CM SNOW

A Duty Forecaster simply checks off any expected events, and verifies the checklist on a following shift. Some weather event, such as a gust to 60 kts, may actually involve three different significant events on the checklist. If winds were forecast of more than 55 kts for instance, three events should have been checked off. If only the 25 and 40 kt categories were checked, then the forecaster would be credited with having forecast two of three significant weather events.

POD scores are simply a reflection of how many events are accurately predicted in advance. If forty of the sixty events were checked off over a particular month, then the POD score would be .67. There is no fine weather bias, as only significant (poor) weather is considered.

False Alarm Ratio (FAR) scores have to balance out the POD scores. An alarmist forecaster may predict lots of significant weather events in every situation, crying wolf at every turn. His POD scores may be high, but this would be balanced out by his poor false alarm ratio. FAR is simply the proportion of all forecasts that are issued which are validated by an event. It is calculated by dividing all accurate forecasts by the total number of forecasts, including false alarms.

Some interesting results were obtained. Table II presents the scores obtained.

For the first month or so, POD scores were very low and FAR scores very high. In November, the first month of the experiment, of 26 events, only 16 were forecast. But of 17 forecasts, 16 were correct. This pointed to the fact that forecasters were far too conservative in forecasting significant weather. Rather than "crying wolf", forecasters adopted the attitude of "wait and see". This had important implications for a weather warning service provided by the weather office. This conservative attitude was overcome to some degree as time went on.

Some other interesting results were obtained. As one can see from Table II, there were marked difference in scores between forecasters. The two extremes were put into Table II, others fell in between these values. POD scores ranged from .40 to .78. Another forecaster had a POD score of .49, but a FAR of .85. In other words, 85% of his forecasts were accurate, but he missed over 50% of the events. It can be deduced that he was too cautious or conservative in his forecasting.

The overall office results showed that of 253 events, 161 were forecast, giving a POD of .64. Maybe this value is low, and the FAR of .81 slightly high. A little more agressive forecasting would probably bring the FAR down to .75, and the POD up into the .70 - .75 range. The difference between the POD and FAR scores represents caution or optimism, something to work on in the future.

The results for the different parameters were also enlightening. Except for snowfall, which had a small sample size, other parameters had similar scores, indicating that the criteria selected represented reasonable values.

CONCLUSIONS

Ranked Probability Score verification methods are reasonable measurements of skill in forecasting. Comparisons with climatology and persistance are good benchmarks. If RPS were to be seriously considered for the CFWS, then it should be done by computer on a routine basis.

POD-FAR scores will be measured at Comox next winter as well. Although a forecaster will receive no credit for skill in forecasting continuing fine weather, he can obtain an objective and simple measure of how he is handling signficant weather, weather that counts. Pessimism and optimism can be detected, and authorities can get some grasp of the margin of safety in the capability to issue weather warnings.

REFERENCES

1. Reid, John D.: Verification of Ceiling and Visibility Forecasts Using the Ranked Probability Score, Atmosphere-Ocean, Vol. 16, No. 2 pp. 177-86, 1978.

2. Charba, P. and Stephen M. Burnham: Comparative Verification of Operational Two to Six Hour Objective Forecasts and Official NWS Watches of Severe Local Storms. Preprint, Conference on Weather Forecasting and Analysis and Aviation Meteorology, October 1978, pp. 156-162.

FIGURE 1

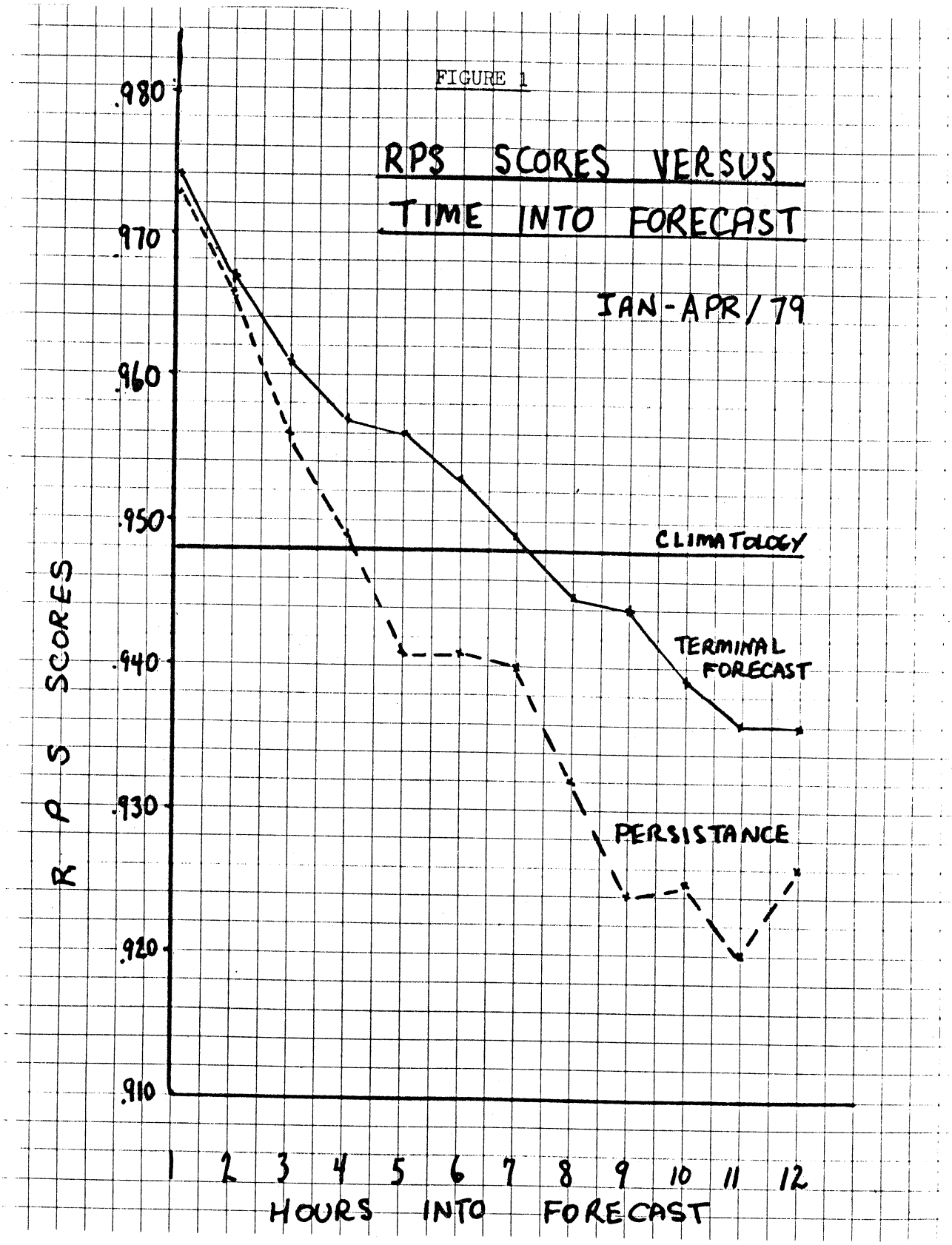
RPS SCORES VERSUS
TIME INTO FORECAST
JAN-APR/79
.980
.970
.960
.950
.940
.930
.920
.910
R P S SCORES
CLIMATOLOGY
TERMINAL FORECAST
PERSISTANCE
1 2 3 4 5 6 7 8 9 10 11 12
HOURS INTO FORECAST

TABLE I

Climatology Forecasts

	Climatological Frequencies (%)						Climatology Forecasts With Limitations					
Weather Cat.	6	5	4	3	2	1	6	5	4	3	2	1
Jan.	58.3	18.8	13.8	2.6	2.0	4.6	0.6	0.3	0.1	0	0	0
Feb.	67.2	15.7	9.9	2.4	1.7	3.1	0.6	0.3	0.1	0	0	0
Mar.	84.3	11.1	3.5	0.8	0.3	0.0	0.9	0.1	0	0	0	0
Apr.	92.3	5.9	1.5	0.3	0.0	0.0	0.9	0.1	0	0	0	0

RPS = .948 RPS = .949

TABLE II

POD and FAR Scores

	No. Events	Cor. Fcsts.	POD	Tot Fcsts	FAR	Skill 1/2[POD + FAR]
Fcstr A	20	8	.40	13	.62	.51
Fcstr B	50	39	.78	46	.85	.81
Fcstr C	35	17	.49	20	.85	.68
Office	253	161	.64	199	.81	.73
Av. Wx.	59	34	.72	46	.74	.73
Wind	139	93	.67	104	.89	.78
Rain	45	31	.69	40	.78	.74
Snow	10	3	.30	9	.33	.32