Precipitation Probability—Comparing Offices for Skill

LAWRENCE A. HUGHES AND WAYNE E. SANGSTER

National Weather Service, Kansas City, MO 64106

(Manuscript received 29 January 1979, in final form 11 June 1979)

ABSTRACT

Using screening regression procedures, an attempt has been made to standardize probability of precipitation Brier scores for difficulty. Climatological factors affecting the difficulty of forecasting used are: precipitation frequency, time persistence and small amount frequency. Standardizing equations were derived for three-month seasons from seven years of data. Four-term regression equations were developed for each season and lead time. Local forecaster improvement over guidance scores varied inversely as the Model Output Statistics (MOS) scores, indicating that poor machine forecasts are easier to improve upon than good machine forecasts.

1. Introduction

Panofsky and Brier (1965) discussed the pitfalls of verification and indicated that one of the greatest dangers lies in attempts to compare relative abilities where there are climatological differences. However, verification is the main way one can get a measure of relative ability, and knowledge of relative ability can be the basis of remedial measures such as training or staffing, and comparing offices can also stimulate interest toward improved forecasts. So we set out to develop standardized scores for comparing offices, even though they are likely to leave something to be desired, at best.

Hughes (1972 and 1979) discussed the effect of four climatological factors on scores of precipitation probability forecasts. These are *precipitation frequency*—Glahn and Jorgensen (1970) showed its effect on Brier (1950) scores, and Hughes, (1968a,b,c, 1969) showed that the effect on the commonly used climatological skill score varied with season, but with little effect in the summer (June–August); *persistence* is significant because it is harder to forecast the starting and stopping of precipitation than to forecast continuation (see Besson, 1924), and this varies with season (Blair, 1924); *trace frequency* is significant because the forecasts are for a measurable amount, so a trace is not a precipitation event. Hughes (1979) showed that trace amounts were the probable cause of major score differences in the winter and fall seasons. *Areal coverage* is significant because the forecasts generally are for the average probability over the local area, and this must equal the observed areal coverage to get reliable forecasts with the best score.

An additional variable that may affect forecaster scores has been introduced in recent years—that of the guidance used by the forecaster. This consists of the Model Output Statistics (MOS) probabilities (Glahn and Lowry, 1972) produced routinely by the National Meteorological Center.

This paper will discuss how the first three of these climatological variables and the quality of the MOS guidance are being allowed for in a comparison of the 14 Weather Service Forecast Offices (WSFO's) of the Central Region (north central portion of the United States) of the National Weather Service (NWS). The areal coverage is not used because data are not available to evaluate it operationally.

2. Data used

Standardizing equations derived by three-month seasons (June–August, etc.) for each of the three forecast periods separately and for each forecast cycle (0000 or 1200 GMT) and based on screening regression procedures were used to compare verification scores. To create the equations, Brier (1950) score differences based on the routine probability forecasts of the 14 Central Region WSFO's and the MOS guidance for them were used. These forecasts were made four times a day at each office, and each forecast consisted of either three contiguous 12 h forecast periods or a 6 h period followed by two 12 h periods. The equations were derived for the two times a day when three 12 h periods were used, and applied without change to the next sets of forecasts (with a 6 h first period), so four forecast sets a day were verified. The available forecasts from January 1972 (when MOS guidance started) through July 1978 were used. Even with some missing data, this gave 16–20 months of dependent data per season per WSFO. Even after separating these data into six sets, with

TABLE 1. List of predictors (all related to 12 h periods).

1. Long-term (15-year) frequency of precipitation and its square.
2. Sample precipitation frequency and its square.
3. The ratio of sample to long-term precipitation frequency and its square.
4. The difference between sample and long-term precipitation frequency, and its square.
5. The long-term conditional probability of precipitation, the conditional period being the 6 h period immediately prior to the first 12 h period forecast.
6. Four types of wet persistence and four of dry persistence based on the same periods as in 5., but using frequencies of the sample.
7. Both the conditional and unconditional probability of precipitation amounts of 0.01–0.09 inch.
8. Brier score (or Brier score squared) of forecasts made using the long-term frequency of precipitation as a probability.
9. Brier score of MOS and its square.
10. IMC = $P_C - P_M$, i.e. the difference in unsquared values of 8. and 9. Predictors 9 and 10 were used only with IFM.

an equation for the day-valid and night-valid forecasts for each of the three forecast periods (six equations each season), there were at least 6500 forecaster probabilities used in deriving each equation.

The predictand data were based on the Brier score of the forecaster's forecasts (P_F), the MOS forecasts (P_M), and climatological forecasts (P_C). These were taken as three differences representing improvements in score as IFM = $P_M - P_F$, IFC = $P_C - P_F$, and IMC = $P_C - P_M$. Thus there were six equations each season for each score. The Brier score was used because it is a "proper" score according to Murphy and Epstein (1967).

The climatological predictor data were obtained from the following: 1) The long-term *precipitation frequency* data (also used to get P_C) were those from Hughes (1966) which were mainly based on 15 years of data up to that time. These are essentially the same as those of Jorgensen (1967); 2) the long-term *persistence* data were from Jorgensen and Klein (1970), where the conditional period was always the 6 h just prior to the first period forecast; 3) no data were available for the *frequency of trace amounts*, so the frequency of small amounts (0.01–0.09 inch), as given by Jorgensen *et al.* (1969), was used instead even though it is probably not completely comparable and thus is less desirable. These basic predictors plus the MOS guidance score and sample relative frequencies and sample persistence were put into a variety of forms as given in Table 1.

3. Results of screening regression

Four-term regression equations were computed; this number was subjectively chosen from experience based on the additional reductions of variance (RV) for each term and the sample size. The total RV's for four predictors are given in Table 2. Note that in all scores the largest average RV was in the winter, with fall next, and spring and summer sharing last.

Also, in all scores the RV was almost universally greatest in the first period of the forecast and except for IFM, least in the third period. IFM standardization had the least RV, with the other two about the same and considerably higher. This tells us that climatic factors, or at least the ones used, have a more significant effect on scores in the winter and first-period forecasts. It also says that the IFM score is relatively free of climatological effects. However, it is believed that generally these effects cannot be neglected in a comparison and ranking (more on this later).

All eight (or 10) basic variables used in a score were picked up by the screening in one period or another as one of the four predictors. However, not necessarily both the square and unsquared parameters were picked up. Since the first predictor selected by the screening almost always contained the bulk of the reduction of variance, let us look mainly at the first predictors.

For the IFC score, the first predictor was mainly (18 of 24 possibilities) predictor 8—the Brier score (or the square thereof) of climatological forecasts, and universally in the sense that the poorer the score of the climatological forecast, the greater the IFC score. This is logical and expected, i.e., IFC should depend heavily but inversely on the quality of the climatological fore-

TABLE 2. Reduction of variance.

Forecast period	IFC				IMC				IFM			
	1	2	3	Avg.	1	2	3	Avg.	1	2	3	Avg.
Winter	0.648	0.488	0.377	0.491	0.607	0.489	0.352	0.483	0.228	0.135	0.177	0.180
Spring	0.458	0.316	0.186	0.320	0.358	0.312	0.223	0.298	0.238	0.129	0.081	0.149
Summer	0.478	0.325	0.224	0.342	0.344	0.305	0.203	0.284	0.254	0.089	0.117	0.153
Fall	0.579	0.434	0.351	0.454	0.515	0.483	0.393	0.463	0.225	0.097	0.162	0.161

cast. The short-term (sample) precipitation frequency and sample persistence were also picked up first on occasion, but only in spring.

The 2nd–4th predictors scattered the possibilities. The conditional frequency of small precipitation amounts (predictor 7) was picked up frequently, but mainly in winter. A high frequency of small amounts of precipitation tended to reduce the IFC score, as expected (Hughes, 1979). Predictors from the persistence group given in Table 1 under predictor 6 were picked most frequently of all, indicating the importance of persistence (dry equally with wet).

The score IMC had essentially the same results as IFC, as might be expected since F and M are very similar forecasts, with both competing against climatology. Thus nothing additional need be said here.

The IFM score is an entirely different matter. The RV's with this score were much less than with the other two. This indicates that this score is much less affected by the climatological factors used, such that predictor 8, so important with the other scores, was of no importance. Relatively little climatological effect would be expected with the IFM score because both the F and the M forecasts depend on the same type of input data and each is exposed to the same climatic conditions. That climatic factors don't drop out completely indicates that MOS is harder to improve upon under some climatic conditions than others. However, in a manner similar to the other two scores, the first predictors selected (but for one) were always 9 and 10, relating to the quality of MOS. Thus even though the RV was sometimes small, the first predictor was related to the quality of MOS, and in the sense that the better MOS scored, the harder it was to improve upon it.

The climatological predictors which were picked up frequently included the abnormality of the precipitation frequency (predictor 4), the frequency of small amounts of precipitation (predictor 7), and again the sample persistence (predictor 6) was prominent. Here the long-term persistence (predictor 5) was more prominent.

The key point here is that no matter what season, or day versus night-valid forecasts, or 1st, 2nd, or 3rd period, the forecaster's ability to improve on *MOS did* depend on the quality of MOS, and inversely, even though this dependence is less than that of IFC or IMC on C, as would be expected. Also it depends to a small extent on climatological factors which can be allowed for.

4. Station ranking

Even after the equations exist to standardize the station's scores for the quality of MOS and for climatological factors, there is still the decision as to what to base the ranking on. The main problems are concerned with what score(s) to use, and if more than one, what weights to give each and then what sample size to use.

As to scores, this question would depend on "the name of the game." If the mission of the forecaster is only to improve on MOS, the IFM score would be sufficient. However, if instead the mission is to get the best forecast to the user, the IFC score should be prime. We used the standardized score of both. This was done by ranking by each of these scores and, for each station, assigning one point for each station below it in rank in each score, adding these points and ranking on this sum as the final ranking. This gives a station credit for a high IFC score. But if it is obtained while making little improvement on MOS, the ranking suffers as a result. The ranking in the IFC and IFM scores has not been the same in any season, so the composite gives a ranking different from that obtained by either score separately.

We started out verifying monthly. However, large shifts in rank were noted from one month to the next so it was decided to use a three-month sample. This seasonal sample (June–August, etc.) is more significant, so changes in rank are of more concern. We also did a four-season ranking with the forecasts through August 1978. The large size of this sample gives its result considerable weight.

The first seasonal application of the normalizing equations was to the fall 1977 forecasts. For those forecasts consisting of three 12 h periods there was usually only a small amount of time between receipt of MOS used in the forecast comparison and forecast issuance (that labeled final guidance), and sometimes this MOS came in after forecast issuance. For the forecasts starting with a 6 h period, the MOS used (that labeled early guidance) *always* came in well after forecast issuance. Having these two sets of forecasts (MOS received before and after forecast issuance) gives additional information on the contribution of MOS as discussed later.

The station ranking in each of the three scores discussed in Section 2 were given to the forecasters in both the unstandardized and standardized versions, each time rankings were made. This let each office see where it ranked in each factor and make judgements therefrom. For example, one office might note in the IFC score that it did very well on a relative basis, i.e., compared to other Central Region WSFO's, but looking at the IFM score, it did not do well. However, one could see that, for example, the reason may have been that MOS did unusually well for their location that time (IMC score). Table 3 gives an example of a seasonal presentation. The (S) indicates a standardized score which is obtained as, for example, IFM(S) = IFM

$-$ IF̂M, where IF̂M is the regression estimate of IFM for the climatological and quality-of-MOS conditions that existed. All scores have been multiplied by 1000. Since the standardized score would vary around zero, 100 is added for psychological reasons. The numbers under total indicate the number of points gained in the standardized ranking in IFC and IFM scores combined, as discussed near the beginning of this section.

The effect of standardization on each score can be easily noted. The largest shift is station G, going up five positions by the standardization of the IFC score. Note that the change in rank from IFM to IFM(S) is a maximum of only two places. However, in the other three seasons of the first year of this program, the standardization of IFM resulted in place changes of four and five in these seasons. Thus standardization of IFM can have a very significant effect on station ranking. The IMC(S) score indicates that MOS had a good season since all but one score is over 100. In spite of this the amount of improvement on MOS by the forecasters was about average since the IFM(S) score had almost as many scores above 100 as below (the average score was 99).

An interesting point, noted in each of the four seasons was found in the IFM score. It had a smaller range but higher average value on the forecasts issued around 0400 and 1600 LST, when MOS is generally available *before* forecast issuance time, than it did for the forecasts issued around 1000–2200 LST, when MOS is received *after* forecast issuance time. This suggests that having up-to-date MOS available for forecast preparation helps the overall score (higher average score) but it hurts the better stations because the best scores were not as high. Thus, as one would expect, MOS helps the weaker stations the most. A mitigating point is that the MOS forecasts compared with those of the forecasters made near 1000 and 2200 LST have the advantage of the more recent LFM (limited-area fine mesh) model data but the forecaster does not have these data available by issuance time. Thus one could also say that having recent LFM data is an advantage.

The ranking based on the whole year of data is the best base for overall conclusions because of its large size. Table 4 gives that ranking in the same style as given under total in Table 3, with the letter for each office unchanged. The Greek letters on the right indicate the type of WSFO. Type α stations have the larger number of experienced forecasters and a smaller turnover of forecasters. Note that all but one of the α stations are in the top six ranks, while the bottom eight are of type β except for one which is a western WSFO (see next). It thus appears that long forecaster experience and a stable staff are definitely conducive to better forecasts.

TABLE 3. Summer (June–August) 1978 ranking—all periods.

IFC		IMC		IFM		IFC(S)		IMC(S)		IFM(S)		Total	
A	59	F	53	B	8	A	126	F	128	A	106	A	26
F	51	A	52	A	7	F	119	A	126	B	106	B	22
E	47	E	46	D	7	E	116	E	122	D	105	C	19
H	40	H	39	C	5	B	113	H	116	C	105	D	18
B	33	K	35	H	2	C	111	K	114	G	102	E	17
C	33	J	29	E	1	H	110	J	114	I	101	F	17
K	30	N	28	G	1	D	109	C	111	H	99	G	15
J	24	C	27	I	1	G	106	B	111	E	98	H	15
N	24	B	25	L	−0	J	105	N	106	F	98	I	11
D	17	M	20	F	−3	K	103	G	106	L	96	J	8
M	16	I	12	M	−4	I	100	M	106	J	95	K	6
I	13	L	12	J	−5	M	98	I	103	K	94	L	4
G	12	G	11	N	−5	N	94	D	103	M	92	M	3
L	12	D	10	K	−5	L	92	L	96	N	90	N	1

Another point can be made here. All of the top six offices are also in the eastern 40% of the Central Region. Thus there may be some climatological or meteorological factor(s) related to predictability or numerical model quality that is not accounted for in the standardization. While not known, this may be because the degree of organization of surface weather systems, and therefore their predictability, increases from west to east across the region. It could also be that the entry of high moisture into the system is more certain in the eastern portion of the region and therefore higher probabilities can be used. These points, if valid, would be difficult to put in numbers. With these two points in mind, it seemed inappropriate to say that skill differences among offices would be wholly given by the objective ranking. This was expected. It was anticipated at the beginning of this ranking effort that only the two or three best and worst offices, as far as forecasting only probability is concerned, should be mentioned as being sufficiently far from the average to justify comment. When the full-year result was sent with comment to the stations involved, the two factors mentioned just above

TABLE 4. Four-season rank.

		Station type
A	24	β
H	23	α
M	22	α
F	21	α
E	18	β
B	15	α
C	13	β
D	10	β
G	9	α
N	8	β
I	7	β
J	5	β
L	4	β
K	3	β

were allowed for because they are items over which the stations have no control. The allowance was subjective, and three stations were mentioned as being higher than one would have expected, while two stations were mentioned as lower than expected. The reasons for these comments were given in terms of the objective ranking and the two subjective factors. One station given credit for above-expected performance was near the middle of the ranking, while one station mentioned as being below expectations was far from the bottom rank. Thus the objective ranking was a guide to station forecast ranking and not the ultimate measure.

5. Conclusions

The quality of station precipitation probability forecasts did depend, as expected, on the climatological factors of precipitation frequency, persistence, and the frequency of small amounts of precipitation. Also, as expected, the ability to do better than some other forecast varied *inversely* with the quality of the other forecast, including MOS. Since the IFM score was comparatively free of climatological effects, it is the best score to use for comparative purposes if these effects cannot be allowed for. However, the quality of MOS should still be allowed for in this case. The relationship can be easily determined by simple regression using forecaster and MOS probabilities of past forecasts.

The Central Region is continuing with this ranking procedure. However, under consideration is a change so only a single score will be used. It would be either IFC, after allowing the multiple regression to consider the effects of variation in MOS quality, or to IFM as currently formulated.

REFERENCES

Besson, L., 1924: Sur la probabilite de la pluie (On the probability of rain). *Comptes Rendus de l'Academie des Sciences*, Paris, **178,** 1743–1745 [English translation, B. M. Varney, 1924: *Mon. Wea. Rev.* **52,** 308].

Blair, T. A., 1924: Local forecast studies—Winter precipitation. *Mon. Wea. Rev.*, **52,** 79–85.

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

Glahn, H. R., and D. L. Jorgensen, 1970: Climatological aspects of the Brier P-score. *Mon. Wea. Rev.*, **98,** 136–141.

——, and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11,** 1203–1211.

Hughes, L. A., 1966: Climatic frequency of precipitation at Central Region stations. ESSA Tech. Memo. WBTM CR-8, 51 pp.

——, 1968a: Seasonal aspects of probability forecasts. 1. Summer. ESSA Tech. Memo. WBTM CR-22, 8 pp. [NTIS PB 185 733].

——, 1968b: Seasonal aspects of probability forecasts. 2. Fall. ESSA Tech. Memo. WBTM CR-23, 15 pp. [NTIS PB 185 734].

——, 1968c: Seasonal aspects of probability forecasts. 3. Winter. ESSA Tech. Memo. WBTM CR-26, 15 pp. [NTIS PB 185 735].

——, 1969: Seasonal aspects of probability forecasts. 4. Spring. ESSA Tech. Memo. WBTM CR-27, 8 pp. [NTIS PB 185 736].

——, 1972: Normalizing precipitation probability verification scores. *Bull. Amer. Meteor. Soc.*, **53,** 72–73 (Abstract).

——, 1979: Precipitation probability forecasts—Problems seen via a comprehensive verification. *Mon. Wea. Rev.*, **107,** 520–524.

Jorgensen, D. L., 1967: Climatological probabilities of precipitation for the conterminous United States. ESSA Tech. Rep. WB-5.

——, and W. H. Klein, 1970: Persistence of precipitation at 108 cities in the conterminous United States. ESSA Tech. Memo. WBTM TDL 31 [NTIS PB 193 599].

——, W. H. Klein and C. F. Roberts, 1969: Conditional probabilities of precipitation amounts in the conterminous United States. ESSA Tech. Memo. WBTM TDL 18 [NTIS PB 183 144].

Murphy, A. H., and E. S. Epstein, 1967: A note on probability forecasting and "hedging." *J. Appl. Meteor.*, **6,** 1002–1004.

Panofsky, H., and G. W. Brier, 1965: *Some Applications of Statistics to Meteorology*, Pennsylvania State University Press, 206 pp.