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

TDL Office Note 76-17

DETERMINATION OF AN OPTIMUM NUMBER OF PREDICTORS FOR PROBABILITY OF PRECIPITATION AMOUNT FORECASTING

Edward A. Zurndorfer and Robert J. Bermowitz

December 1976

by

Edward A. Zurndorfer and Robert J. Bermowitz

INTRODUCTION

Since February 1975, objective forecasts of the probability of precipitation amount (PoPA) and categorical forecasts of precipitation amount have been supplied as guidance to the Quantitative Precipitation Branch of the National Meteorological Center (NMC) (Bermowitz and Zurndorfer, 1975). With use of the Model Output Statistics (MOS) technique (Glahn and Lowry, 1972), we develop 12-term regression equations for the categories $\geq .25$, $\geq .50$, ≥ 1.0, and ≥ 2.0 inches for various projections from both the 0000 GMT and 1200 GMT cycles.

Although our predictands may be correlated with many variables, a regression equation containing only a few may explain about as much of the variance as an equation containing many. This is the result of high intercorrelations among the variables. Also, even though an equation containing many predictors may estimate the predictand rather well on the dependent data, it may be showing not only the real physical relationships but also the chance relationships that will not be present in independent data samples. Therefore, the equation with many terms will not necessarily perform as well on independent data as one with fewer terms. We wanted to find out whether we should continue to use 12-term equations.

In addition to investigating the optimum number of predictors to use in our regression equations, we also looked at some other questions concerning PoPA forecasting, including:

(1) Does mixing continuous predictors with binary predictors give better results than using binaries alone? Continuous predictors were found to be useful in probability of precipitation (PoP) forecasting (Glahn and Bocchieri, 1976).

(2) Does the use of an arbitrary cutoff reduction of variance criteria in our screening procedure improve the PoPA forecasts when compared to stopping the screening procedure at a particular number of terms?

This paper describes our efforts to answer these questions for the purpose of improving our PoPA forecasts.

TEST PROCEDURE

We developed variable term equations for the 12-36 hr projection after 0000 GMT using five cool seasons (October-March, 1970-71 through 1974-75) of beyond 14 terms because other attempts to determine the optimum number of predictors (Bocchieri and Glahn, 1972) have shown that little improvement is achieved beyond 14 terms. Also, we developed 10-, 12-, and 14-term equations with both binary and continuous predictors after initial testing on the binary equations indicated that this was about the optimum number of predictors to use.

For each type of equation, we developed generalized operator equations for the categories $\geq .25$, $\geq .50$, and ≥ 1.0 inch for each of the 9 regions shown in Figure 1. These 9 cool season regions are the same ones used in the current operational PoPA system (Bermowitz and Zurndorfer, 1975). As was the case for the operational system, equations for all categories were derived simultaneously, so that the chosen predictors are used in all equations. This procedure provides for greater consistency among the probability forecasts of the different categories.

Predictors screened in developing the equations included forecast fields from the primitive equation (PE) (Shuman and Hovermale, 1968) and the trajectory (TJ) (Reap, 1972) models. Continuous predictors screened in the development of binary and continuous equations were determined from a list of "best" binary predictors used in the development of binary predictor equations. We also screened sine and cosine day of the year and station elevation for all types of equations.

We also developed equations using a cutoff reduction of variance (RV) of .005. We chose .005 because we wanted a value that would ensure that each equation developed would contain at least 2 or 3 terms. The number of terms in these equations ranged from 3 for the driest regions to 9 for the wettest region; the average number for all regions was 7.

To test the relative performance of each type of equation, we computed P-Scores (Brier, 1950) for all 3 categories and expressed these scores as improvements over climatology for both the development and an independent data set. As pointed out by Glahn and Jorgensen (1969), expressing P-Scores as a function of climatology is useful in comparing different sets of forecasts.

RESULTS AND CONCLUSIONS

The following results are shown in Tables 1 and 2:

(1) For equations containing only binary predictors and also for those containing both binary and continuous predictors, 14-term equations had, of course, the highest improvement over climatology on the dependent data. However, the differences between 10-, 12-, and 14-term equations were small. On the independent data, the improvements over climatology averaged about the same for the 3 categories for the 12- and 14-term equations.

12-, and 14-term equations), improvements over climatology increased. The equations containing continuous predictors "held up" better on independent data than did the equations with only binary predictors.

(3) Use of a cutoff RV of .005 did not improve the PoPA forecasts. In fact, this criteria produced results somewhat comparable to those obtained with the use of a 6- or 8-term binary equation. This corresponds to the average number of terms (7) used in the equations with a .005 cutoff RV.

The results indicate that 12 or 14 terms are the optimum number to use in our PoPA equations and that these equations should contain both binary and continuous predictors. We feel, however, that with 12 predictors we have likely reached the "noise level"; that is, we have reached the point where the small additional reductions of variance are due mainly to chance relationships between predictors and predictand. Until now, we have been using 12 terms in our equations and, as a consequence of the results shown here, we plan to continue using this number in the future. We also plan to screen continuous predictors in addition to our usual binaries.

It should be remembered that these experiments were performed on cool season data. However, we feel that the results with warm season data (April-September) would be similar. Also, we used 2 models, the PE and TJ, in performing these experiments; perhaps the use of a 3rd model such as the Limited Area Fine Mesh model (LFM) (Howcroft and Desmaris, 1971) and surface observations in combination with the other two would yield slightly different results.

REFERENCES

Bermowitz, R. J., and E. A. Zurndorfer, 1975: Current status of probability of precipitation amount (PoPA) forecasting. TDL Office Note 75-10, 6 pp.

Bocchieri, J. R., and H. R. Glahn, 1972: Use of model output statistics for predicting ceiling height. Monthly Weather Review, 100, 869-879.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78, 1-3.

Glahn, H. R. and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. Journal of Applied Meteor., 11, 1203-1211.

_____, and D. J. Jorgensen, 1969: Climatological aspects of the Brier P-Score. Monthly Weather Review, 98, 136-141.

Review, 104, 127-132.

Howcroft, J. and A. Desmaris, 1971: The limited area fine mesh model. NWS Technical Procedures Bulletin, No. 67, 11 pp.

Reap, R. M., 1972: An operational three-dimensional trajectory model. Journal of Applied Meteor., 11, 1193-1202.

Shuman, F. G., and J. B. Hovermale, 1968: An operational six-layer primitive equation model. Journal of Applied Meteor., 7, 525-547.

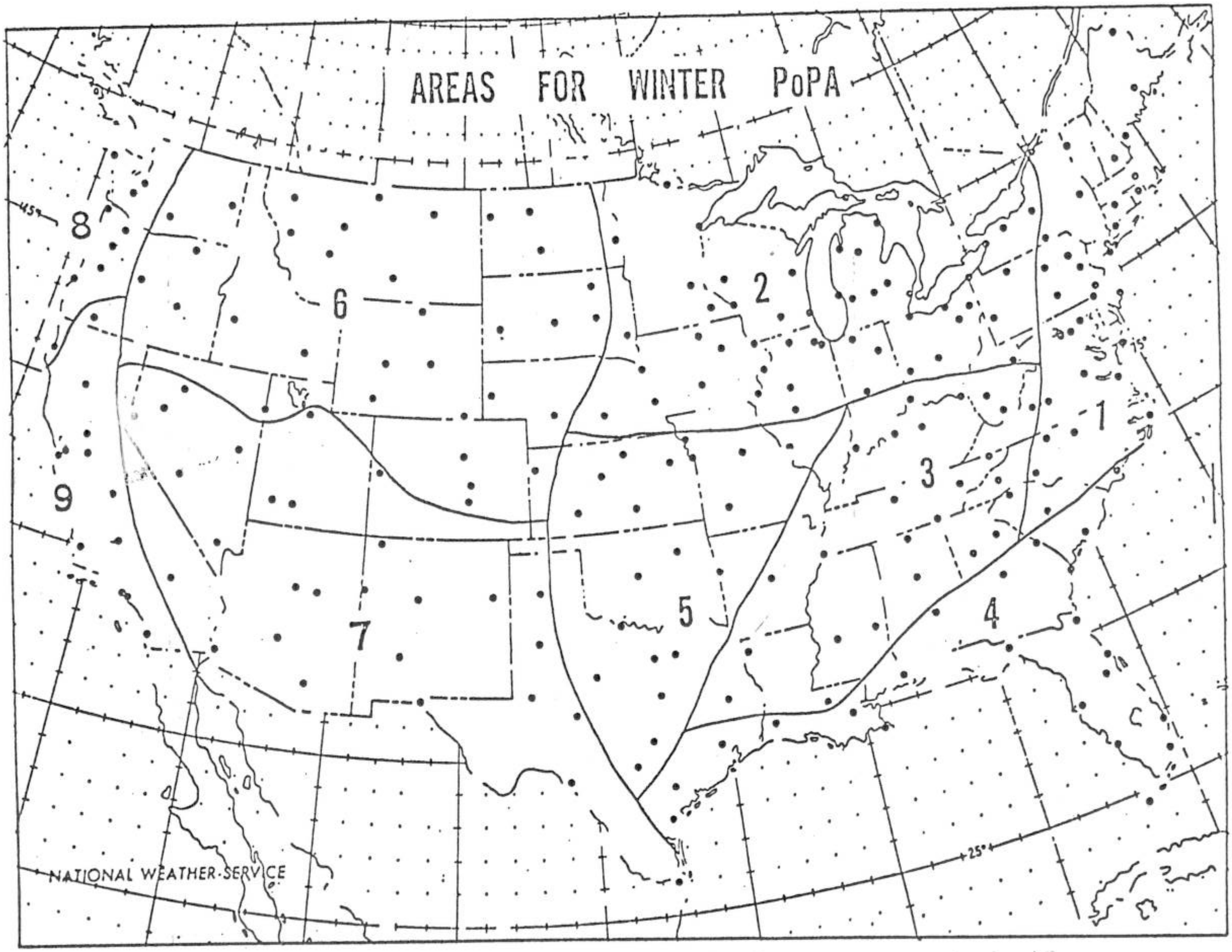

Figure 1. The 9 regions used for developing equations.

the 737-day dependent data sample.

Number of Terms in the Equations	Type of Predictors Used	Improvement over Climatology for Categories		
		≥.25	≥.50	≥1.00
2	Binary	23.26	16.36	5.85
4	Binary	26.67	19.10	8.24
6	Binary	28.04	20.43	9.10
8	Binary	28.85	21.10	9.67
10	Binary	29.36	21.55	10.23
12	Binary	29.78	21.85	10.44
14	Binary	30.05	21.98	10.73
10	Binary & Continuous	30.18	22.70	11.51
12	Binary & Continuous	30.41	22.86	11.80
14	Binary & Continuous	30.55	22.98	11.98
Cutoff Reduction of Variance of .005	Binary	28.66	20.99	9.67

the 176-day independent data sample from the 1975-76 cool season.

Number of Terms in the Equations	Type of Predictors Used	Improvement over Climatology for Categories		
		≥.25	≥.50	≥1.00
2	Binary	22.86	15.58	4.90
4	Binary	24.73	17.00	5.61
6	Binary	26.40	18.61	6.36
8	Binary	27.30	19.19	6.72
10	Binary	27.66	19.65	7.19
12	Binary	28.02	20.03	7.47
14	Binary	28.28	20.05	7.11
10	Binary & Continuous	29.06	20.27	9.12
12	Binary & Continuous	29.24	21.27	9.22
14	Binary & Continuous	29.25	21.23	9.08
Cutoff Reduction of Variance of .005	Binary	27.14	19.33	6.94