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

TDL Office Note 78-14

A MODEL FOR CONVERTING PROBABILITY FORECASTS TO CATEGORICAL FORECASTS

Robert G. Miller and Donald L. Best

December 1978

A MODEL FOR CONVERTING PROBABILITY FORECASTS TO CATEGORICAL FORECASTS

Robert G. Miller and Donald L. Best

1. INTRODUCTION

This Office Note introduces a general purpose threshold probability model that provides a means to convert probability forecasts into a wide range of categorical responses. A threshold probability is a number between zero and one used as a cutoff in deciding which event to forecast categorically, given a set of forecast probabilities. The need for threshold probabilities is very basic: How does one convert a forecast probability distribution into a single decision that is favorable, or in some way optimum, for a given customer's operation? In decision theory, this can be done by applying a utility function to the forecast probability distribution. To many users of weather information, however, the determination of a utility function is not easy. Precise threshold probabilities can substitute for utility functions as long as the categorical forecasts match acceptable decision frequencies, but they are likewise not easily derived. The threshold probability model described herein can, however, cover most customer utility needs through the selection of a preferred set of categorical decision frequencies.

2. HISTORICAL CONSIDERATIONS

Thompson (1952) was one of the first meteorologists to tackle the problem of determining threshold probabilities. In a simple two-event situation it was the ratio of the cost to protect against an adverse event over the loss incurred should the event occur without protection.

In the two or more event problems, the classical decision theoretic approach chooses an optimum decision using Bayes' Rule which minimizes the expected loss (see Miller 1962). A utility function expressing the loss or gain for all possible outcomes is required, while threshold probabilities are only used implicitly.

Many people tried to estimate general utility functions such as Allen (see Glahn, 1964), Gringorten (1967), and Bryan and Enger (1967). None of these really achieved popularity. The problem is that weather forecasts have a wide spectrum of users and no single function satisfies them all. Murphy (1974) made a concerted effort to deal with the situation but did not go beyond the three group problem.

Klein, Lewis, and Enger (1959) approached the problem by attempting to derive threshold probabilities through use of statistical parameters such as climatology and correlation, the latter being a measure of the goodness of forecast fit. This has worked well in the two event situation. For the more than two event case Bermowitz (1978) produced a systematic empirical approach which is in current operational use by TDL. The method, however is quite lengthy and uses much computer time and manpower.

These research efforts have not solved the problem of having a general threshold probability model that has all things for all users nor are any current solutions necessarily simple or easy to use. The general threshold probability model described in this Office Note is the most recent attempt to reach that goal.

3. DESCRIPTION OF THE MODEL

A. General Model

The form of the general threshold probability model, hereafter referred to as the M&B model, is

$$p^* = FR(.5 - C) + C$$

where p* is the computed threshold probability value given the relative frequency, or climatology, (C) of the predictand event, the multiple correlation coefficient (R) of the forecast equation, and an adjustment term (F) which can range between 0 and R^{-1} according to desired verification effect. The M&B model can be best understood by examining some of its unique variations as depicted through user-desired statistics such as chi-square minimization and threat score maximization or by its extreme solutions which replicate the maximum probability and the climatology threshold models described below.

B. Maximum Probability (Max Prob) Model

The max prob model is p*=0.5. The M&B model reduces to this form at F=1 and R=1. The constant threshold value of one-half is equivalent to a decision rule whereby the category with the larger probability is selected. This model is perhaps the one most commonly used because of its simplicity and desirable feature of producing the highest percent of correct forecasts. On the other hand, its primary weakness is its general inability to forecast the rarer events. For example, a forecast greater than P=.5 is very unlikely to be produced by a forecast equation whose relative frequency of predictand event is only C=.05. Hence, for the rarer event the max prob model will not be useful. This problem can be of deep concern to some users of weather information. A model which overcomes this deficiency by design is the climatology, or Gringorten model.

C. Climatology (Gringorten) Model

The Gringorten model is p*=C. The M&B model reduces to this special case at R=0. Gringorten (1967) recognized the problem with the max prob model and introduced the use of relative frequency of the event, or climatology, as a threshold model. This model, however, has a similar deficiency to the max prob model but in the opposite sense. That is, the rarer event category is overforecast. This may be considered an ideal performance by some customers concerned with damaging events such as severe thunderstorms or heavy precipitation. To others, however, following such "cry wolf" forecasts could dilute the usefulness of this model to the real economic problem at hand. There is also need for

a middle ground response such as a balance between the frequency of forecast and observed events. A term commonly used to denote such balance is unit bias[1].

D. Unit Bias Model

The unit bias model is $p^*=R(.5-C)+C$. This is the M&B model over the range of R at F=1. The unit bias model is designed to forecast the rarer event as often as it occurs and is found in the solution space between the max prob and Gringorten extremes (i.e. between R=0 and R=1) as illustrated in Figure 1. A special case also shown in Figure 1 is the M&B solution for maximizing the threat score[2] statistic.

E. Maximum Threat Model

The max threat model is $p^*=.698R(.5-C)+C$. This is the M&B model over the range of R at F=.698. The use of this model permits a certain degree of overforecasting the rarer event in such a way as to maximize the threat score and yet stay somewhat away from the "cry wolf" problem. The F-value of .698 was derived from known relationships among climatology (C), forecast equation performance (as measured through R), and optimum threshold probability thresholds for maximum threat results. These data were provided by Bermowitz (1978) from his work with equations which forecast precipitation amounts. The solution of F=.698 was derived by linear regression and gives this special case of the M&B model a 0.943 reduction of variance within the data sample.

4. VERIFICATION

Several experiments on dependent data samples were made in the course of the development work on the M&B model. For purposes of illustration, however, only two data tests are shown in this Office Note. To support the claim that the M&B model encompasses many features, the tests are on independent data. The categorical selection procedure used is a pairwise comparison test. That is, the forecast probability P will always refer to the category or combination of categories under decision consideration. In the case where multiple predictand group probabilities are considered, the pairwise procedure is applied as follows: (1) compare the probability of the first category against the first threshold probability and decide "yes or no." (2) If "yes," category one is picked as the best categorical forecast and the process stops. (3) If "no," add the probabilities for both category one and two and compare the new result to the next threshold probability value. (4) If "yes," category

---

[1] Bias is defined here as the fraction F/O where F and O are the number of forecasts and observations of the same category, respectively. Unit bias is attained when F and O are equal.

[2] Threat score is defined as H/(F+O-H) where H is the number of correct forecasts of the rarer event category. F and O are defined above.

two is selected and the process stops. (5) If still "no", proceed by including the third category probability with the previous two groups and continue testing. This pairwise selection procedure will select the last category by default.

5. RESULTS

A. Visibility Test

The forecast probabilities used to produce Table 1 are from an 18-h Alaskan visibility MOS equation set. The independent test period is March-May 1977. This is a generalized operator equation set developed on a sample composed of surface observations, special constants, and interpolated LFM fields at 14 Alaskan stations. All four variations of the M&B model were tested with pertinent statistics listed in Table 1. For simplicity, a dichotomous forecast situation is chosen (i.e. forecasting visibility below or above six miles--note the relative frequency of less than six miles is 20%). As expected, the max prob model gained the highest percent of correct forecasts and also underforecast the rare event (note the 0.49 bias as compared to the others each above 1.0). The unit bias model produced the best forecast-to-observed frequency fit as verified by its lowest chi-square statistic of 1.1. The critical chi-square value for this one degree of freedom test is 3.84 at the 5% level. Notice that the other models failed to produce a non-significant chi-square. The minimum chi-square score is also reflected through the biases which are closer to 1.0 than the other models. The Gringorten model also performed as expected in that it produced significant overforecasting of the rarer event to the detriment of the percent correct and the skill score statistics. This overforecasting also did little to improve the threat score over the unit bias model. The max threat model performed as expected as shown through the threat score statistic being largest for this model. It is also worthy of note that this model produced the best biases after the unit bias and max threat models. The results in Table 1 show the flexibility of the M&B model to produce a range of prespecified results.

B. Sky Cover Test

The forecast probabilities used to produce Table 2 are for 18-h Alaskan sky cover. The MOS equations are single station and verified on an independent data sample for the period March-May 1977. The sky cover for this test was categorized into two groups: "clear" and "not-clear". Table 2 is comparable to Table 1 in the sense that the same sample size, sample statistics, and Alaskan weather were considered. However, there is one significant difference to note: Table 1 is derived from a single forecast equation set that is valid for all 14 stations, whereas Table 2 comes from 14 separate equation sets being applied to the sample. Table 2 represents, therefore, a compaction of 14 separate equation sets performances. The results are all expected except for one--the max threat model did not produce the maximum threat score. Although it produced a high treat score of .470, the unit bias model

performed better at .484. Although this result could be dismissed as sample fluctuations, a special investigation was undertaken to determine if there was a more descriptive reason. This apparent failure suggests perhaps that the value of F could be inadequate for this particular sample. The F value was derived from a data sample where the input parameters were limited to $.079 \leq R \leq .573$, $.002 \leq C \leq .161$, and $.020 \leq p^* \leq .280$ with sample size of 166. To analyze this possibility a scatter plot (Fig. 2) of the 14 station's (R,C) pairs were overlayed against the development model's (R,C) data ranges mentioned above. Since only three station's (R,C) pairs lie within these boundaries, it appears that our suspicion about the value of F may be correct. To verify this, however, another analysis of only those three stations (Table 3) was accomplished. Table 3 (although produced from a much smaller sample size) shows that the max threat model did produce the highest threat score when the (R,C) pairs were in the same range as those on which the F=.698 was determined.

6. SUMMARY AND CONCLUSIONS

TDL presently uses an iterative procedure (Crisci, 1976; Glahn, 1974) on the dependent data sample to provide operational MOS programs with appropriate product-related threshold values. For example, ceiling and visibility requires threshold values that will give their categorical forecasts a unit bias response while precipitation amounts and thunderstorm categories need to be selected with the maximum threat score statistic in mind. Use of the M&B model to produce the required threshold values instead of the current iterative approach will save many man-hours and computer time in the development of new or rederived MOS forecast equations. This is accomplished by using the M&B model in a subroutine to the regression equation software and running the subroutine as a final step.

The results of the two experiments shown in Tables 1 and 2 illustrate that the M&B model is truly responsive to various threshold probability needs--hence, customer's categorical decision needs.

For a customer of weather information, particularly when a categorical decision must be made, the choice of actions ideally should be tailored to his needs. If these four model's characteristics were shown to the customer beforehand, utility could be approached in a different manner; that is, establish the customer's desired forecast performance as closely as possible to one of the four models. For example, if maximum hits is desired, select the max prob model; if best balance between forecast and observed frequencies is desired, select the unit bias model; or if well balanced overforecasting is needed as measured by the threat score statistic, select the max threat model. Whereas this has been described as a selection of one of the four given models, there is no limit to the possible number of models the user can choose from.

REFERENCES

Bermowitz, R. L., 1978: Personal Communication.

_____, and E. Zurndorfer, 1978: Automated Guidance for predicting quantitative precipitation. Mon. Wea. Rev., (In press).

Bryan, J. G., and I. Enger, 1967: Use of probability forecasts to maximize various skill scores. J. Appl. Meteor., 6, 762-769.

Crisci, R. L., 1976: Improving the bias in MOS ceiling and visibility forecasts. TDL Office Note 76-4, National Weather Service, NOAA, U.S. Department of Commerce, 9 pp.

Glahn, H. R., 1964: The use of decision theory in meteorology with application to aviation weather. Mon. Wea. Rev., 92, 32-35.

_____, 1974: Problems in the use of probability forecasts, Preprints Fifth Conference on Weather Forecasting and Analysis. Amer. Meteor. Soc., Boston, Mass., 4 pp.

Gringorten, I. I., 1967: Verification to determine and measure forecasting skill. J. Appl. Meteor., 6, 742-747.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. J. Appl. Meteor., 16, 672-682.

Miller, R. G., 1962: Statistical prediction by discriminant analysis. Meteor. Mono., 4, Am. Meteor. Soc., Boston, Mass., 54 pp.

Murphy, A. H., 1974: Evaluation of probability forecasts in meteorology. Unpublished Doctoral Thesis. University of Michigan.

Thompson, J. C., 1952: On the operational deficiencies in categorical weather forecasts, Bull. Am. Meteor., Soc., 33, 223-226.

Table 1. Alaska 18-h Visibility MOS. Generalized operator equation (valid for all 14 locations) with an independent sample (March-May 1977) of 1191 cases. Selected verification statistics for four threshold probabilities of the basic M&B model. Underlining signifies best result per column.

Model	Biases		Pcnt Hits	Skill	Chi-Sq	Threat
	< 6 mi	≥ 6 mi				
Max Prob	.49	1.15	<u>79.3</u>	.268	28.5	.231
Unit Bias	<u>1.12</u>	<u>.97</u>	75.6	.324	<u>1.1</u>	.319
Gringorten	2.59	.54	58.7	.239	156.	.320
Max Threat	1.38	.89	73.5	<u>.327</u>	11.2	<u>.335</u>
Rel. Freq. (Dependent)	.20	.80				

Note: Chi-square test on 1-DOF at the 5% level is 3.84.

Table 2. Alaska 18-h Sky Cover MOS. Single station equations (one equation per station) with an independent sample (March-May 1977) of 1191 cases. Underlining signifies best result per column.

Model	Biases		Pcnt Hits	Skill	Chi-Sq	Threat
	Clr	Not Clr				
Max Prob	.89	1.03	<u>85.6</u>	.534	11.3	.452
Unit Bias	<u>1.10</u>	<u>.98</u>	85.3	<u>.559</u>	<u>6.77</u>	<u>.484</u>
Gringorten	3.24	.43	34.5	-.107	625.	.132
Max Threat	1.25	.94	83.6	.535	10.7	.470
Rel. Freq. (Dependent)	.23	.77				

Note: Chi-square test on 14-DOF at the 5% level is 23.7.

Table 3. Alaska 18-h Sky Cover MOS. Three station subset of Table 2 results consisting of 235 cases. Underlining signifies best result per column.

Model	Biases		Pcnt	Skill	Chi-Sq	Threat
	Clr	Not Clr	Hits			
Max Prob	.00	1.05	<u>95.7</u>	.000	5.73	.000
Unit Bias	<u>.64</u>	<u>1.02</u>	94.0	<u>.193</u>	<u>1.14</u>	.125
Gringorten	19.9	.07	10.6	-.002	190.	.045
Max Threat	<u>1.36</u>	<u>.98</u>	91.5	.187	2.57	<u>.130</u>
Rel. Freq. (Dependent)	.06	.94				

Note: Chi-square test on 3-DOF at the 5% level is 7.81

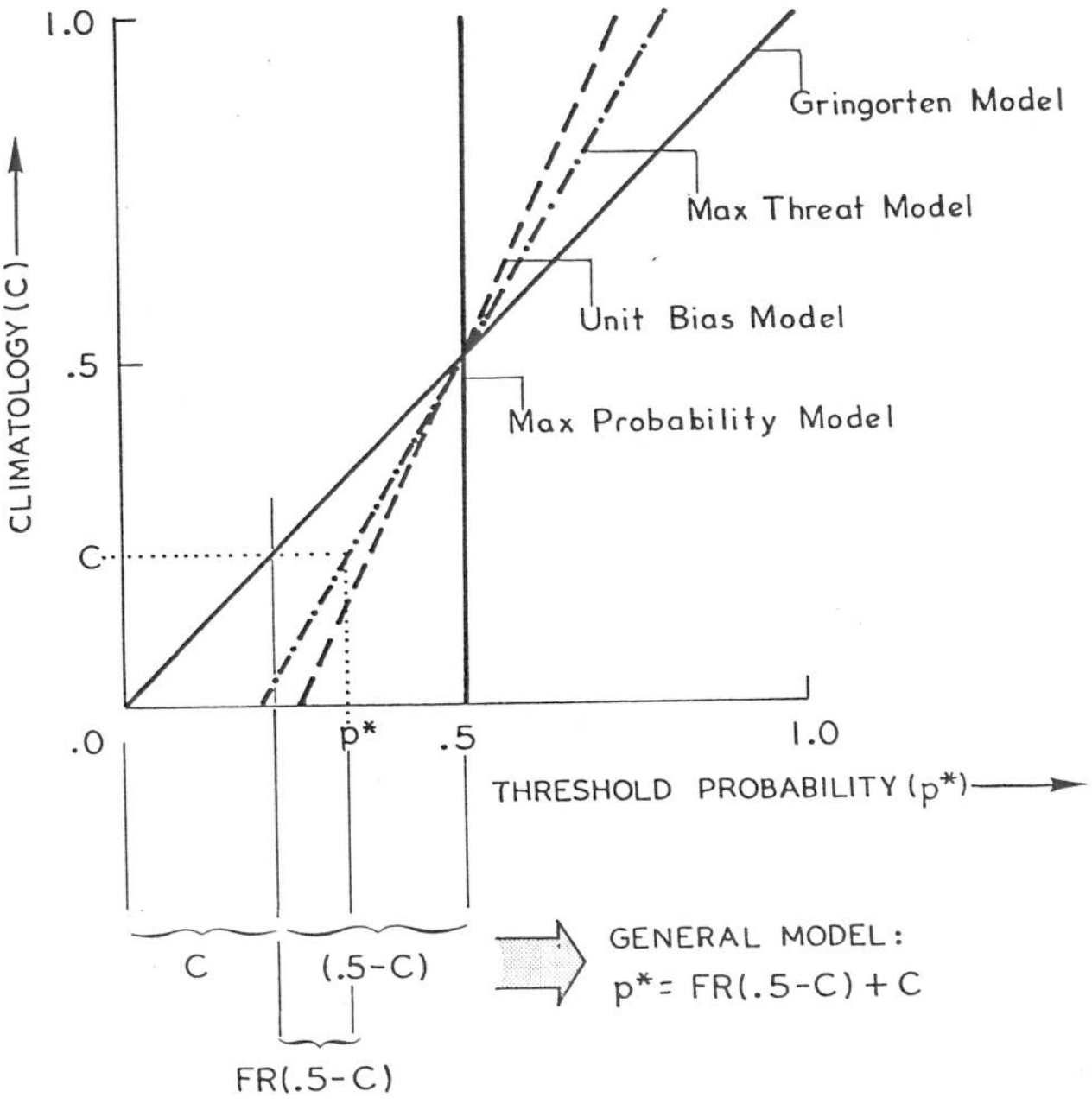

Figure 1. Relationship among various threshold probability models and the solution space of the general model.

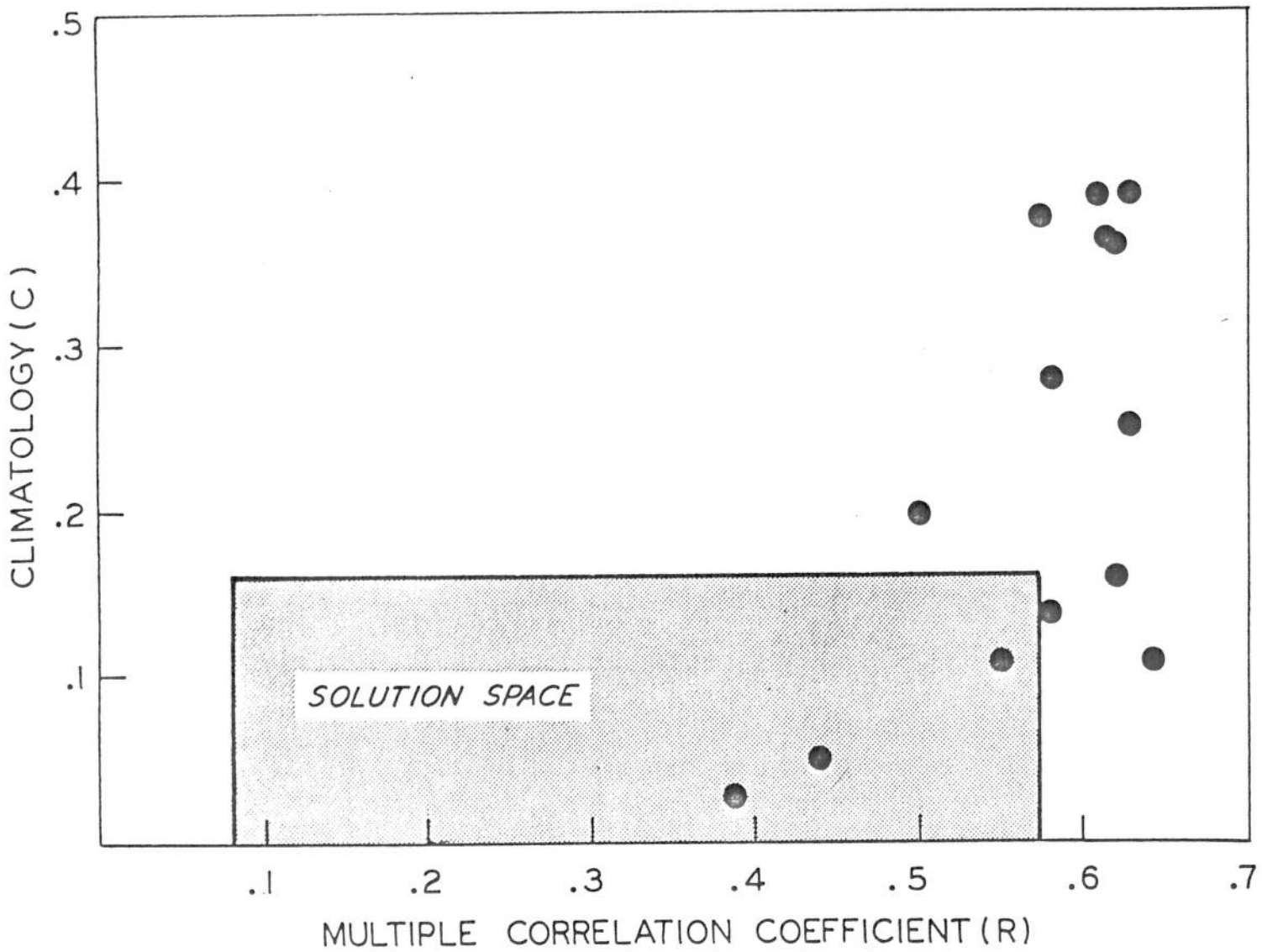

Figure 2. Comparison of individual Alaskan station's (R, C) pairs to the solution space of the (R, C) pairs used to compute maximum threat weighting factor F = .698.