Forecast Skill Scores

IMPROVING YOUR WEATHER FORECASTS THROUGH A BETTER KNOWLEDGE OF SKILL SCORES

Robert L. Vislocky (1)
George S. Young (2)
Pennsylvania State University
University Park, PA 16802

ABSTRACT

Several skill scores that are used to evaluate the quality of weather forecasts will be discussed. Each skill score can be optimized by following the associated directive when making a forecast. If a forecaster wishes to do well with respect to a certain skill score, he owes it to himself to understand and follow the appropriate directive. Knowledge of these directives can be applied to forecast contests, issuing public weather forecasts, and understanding objective forecast systems.

1. INTRODUCTION

Successful forecasting of surface weather elements (cloud amount, temperature etc. . . .) requires two components. One is a comprehensive knowledge of what meteorological events could happen in the forecast period. The other is a statistical knowledge of how to translate the event probabilities into a weather forecast in such a way that the forecaster can optimize a skill score. Much attention has been devoted to improving the meteorological component of forecasting, while relatively little attention has been given to the statistical component. This paper will address the latter topic.

Many different skill scores have been used to evaluate the quality of weather forecasts. Associated with each skill score is a directive which a forecaster must follow or if he or she is to optimize that skill score. A directive is an objective method of converting the predicted probability distribution for a weather element into a discrete number or category. For an example of a directive associated with a given skill score, consider the situation a professional gambler must face when he is selecting a horse to bet on in a race. Suppose there are four horses in the race and the gambler assesses each horse's probability of winning as follows (assume the gambler's probabilities are reliable and the payoff odds are not influenced by the gamblers' bet):

HORSE #	WIN PROBABILITY	PAYOFF ODDS
1	50%	50% (1-1)
2	20%	20% (4-1)
3	20%	10% (9-1)
4	10%	20% (4-1)

Which horse should the gambler bet on? Obviously horse #1 has the highest probability of winning. Betting on the horse that has the highest probability of winning is a directive that is consistent with maximizing the percent correct score. However, horse race betting is not evaluated by the percent correct score! The score to maximize in horse race betting is money earned. The directive to follow when maximizing money earned is to bet on the horse that has the highest win probability to payoff odds ratio. Therefore, the gambler should bet on horse #3 since a bet on horse #3 is consistent with the directive that maximizes money earned.

Thus, by example, we can see that a forecaster must take into account the score by which his forecasts are being evaluated. In addition, the forecaster must follow the appropriate directive if he is to optimize that particular score. In the rest of this paper, several skill scores which are frequently used to evaluate the quality of weather forecasts will be examined and the directives to be followed when trying to optimize these skill scores will be discussed.

2. [FORECAST-OBSERVED] SCORE

The directive to follow for minimization of [forecast-observed] error points is to forecast the median event of the predicted probability distribution (3). As an example, consider a situation which arises frequently in the National Collegiate Weather Forecast Contest. In this contest, precipitation is broken down into the six categories given below.

CATEGORY	AMOUNT
0	0″
1	.01–.05″
2	.06–.24″
3	.25–.49″
4	.50–.99″
5	> .99″

Imagine a day where the probability of .01″ or more of rain is ony 40%. However, if it does rain, then there is a good chance of having greater than .25″. This situation happens on many days where spotty, deep convection occurs. The forecasted probability distribution for this scenario might look like this:

CATEGORY	0	1	2	3	4	5
PROBABILITY	60%	5%	10%	10%	10%	5%

How much rain should be forecast? Obviously there is a good chance of heavy rain. However, the directive to follow is to forecast the median event, which is category 0. This is the only category that can minimize the [forecast-observed] error points in the long run. To see this, consider 20 forecasts made under meteorological conditions similar to the situation above. The relative frequencies of occurrence of each category are given below:

CATEGORY	PROBABILITY	FREQUENCY
0	60%	12/20
1	5%	1/20
2	10%	2/20
3	10%	2/20
4	10%	2/20
5	5%	1/20

If a forecaster predicts category 2 in an attempt to hedge between heavy precipitation occurrences and no precipitation occurrences, then the accumulated error points will be

34. One error point is totaled for each category by which a forecast is off. If a forecaster predicts the median event, category 0, his error points will only total 24. This represents a 29% improvement in skill.

3. (FORECAST-OBSERVED)^2

This is a score which is frequently used to evaluate probability forecasts. The directive to follow for minimization of (forecast-observed)^2 error points is to forecast the mean of the predicted probability distribution (3). As an example, consider a contest where the object is to forecast the probability of .01" or more of rain at a certain station. A binary system is used so that if the event does occur then the observed value equals one and if the event does not occur then the observed value equals zero. The forecast value represents the probability the forecaster places on the likelihood of occurrence of the event. The probability can be any value between zero and one, inclusive.

Imagine a situation where a storm is moving over the station for which you are forecasting. Let's say that this storm has only a 50% areal coverage of precipitation and that the precipitation is randomly distributed around the forecast station. Then we can say that precipitation will occur at that station on average 10 times out of 20 forecasts. Using the binary system we should forecast (10(1) + 10(0))/20 = .50 because this is the mean of the event distribution. A forecast of any value other than .50 will, in the long run, ruin one's (forecast-observed)^2 score. To see this, consider 20 similar cases. If a forecaster predicts 70% in an effort to try to hit the rain events, his accumulated error points will be $10(1-.70)^2 + 10(0-.70)^2 = 5.8$. If the forecaster follows the correct directive and forecasts .50 each time then his accumulated error points will only be $10(1-.5)^2 + 10(0-.5)^2 = 5.0$. This represents a 14% improvement.

Frequently in a case like this, an inexperienced forecaster in a forecast contest will predict something other than 50%. Experience at the Pennsylvania State University shows that many forecasters absolutely refuse to forecast 50% probabilities because they feel a such forecast represents a cop out. In addition, they prefer forecasting sharper probabilities because they feel sharper forecasts have a higher utility. This sharpening of forecasts beyond the objective probability of occurrence hurts their skill scores in this type of evaluation and, in the long run, decreases utility.

To further illustrate the importance of following the appropriate directive when making a forecast, it is useful to evaluate the results of a probability forecast contest (4). Sanders reported that "consensus" forecasting outperformed most, if not all, of the individual forecasters in the contest. A consensus forecast is the mean of all of the individual forecasts. By taking the mean of all the individual forecasts, consensus is following the directive of the (forecast-observed)^2 score! This is a major reason why consensus performs so well in probability contests.

4. PERCENT CORRECT AND THREAT SCORES

Many times weather forecasts are categorical (yes it will rain or no it won't rain for example). This is especially the case in many television and radio forecasts of precipitation where there is a desire not to "clutter up the forecast with probability numbers." The question then arises as to the minimum probability that is required before one issues a forecast of rain. Should one issue a forecast of rain only when the point probability exceeds 50%, or should the threshold be lower, say 30%? A low threshold probability means the forecaster will hit more rain events but will also have a high false alarm rate. To begin to answer this question consider the following contingency table:

		Forecast	
		Yes	No
Observed	Yes	A	B
	No	C	D

Box A represents the number of times a forecaster predicted rain and rain was observed (# of hits). Box B represents the number of times a forecaster predicted no rain and rain was observed. Box C represents the number of times a forecaster predicted rain and no rain fell (# of false alarms). Box D represents the number of times a forecaster predicted no rain and no rain fell. The percent correct score is (A + D)/(A + B + C + D) and the threat score is A/(A + B + C).

To show the dependence of these scores on the threshold probability one chooses, data from a precipitation probability experiment (5) will be used. In this experiment, probability forecasts given by National Weather Service forecasters were converted into yes/no rain forecasts depending on whether or not that probability exceeded various thresholds. The values for each contingency table element are shown below for varying threshold probabilities.

THRESHOLD	A	B	C	D
60%	144	628	68	3540
50%	233	539	163	3445
40%	346	426	317	3291
30%	515	257	685	2923
20%	660	112	1387	2221

The main inference one can make from this table is that as the threshold probability lowers, the number of hits (element A) increases and the number of false alarms (element C) also increases. The table below shows the effect of the threshold probability on the percent correct and threat scores.

THRESHOLD	% CORRECT	THREAT
60%	84.1	.171
50%	84.0	.249
40%	83.0	.318
30%	78.5	.353
20%	65.8	.306

From this table we see that as the threshold lowers (as the forecasts get wetter) the threat score first increases towards the maximum at the 30% threshold, and then decreases. The percent correct score decreases dramatically as the threshold is decreased from 50% to 20%. From these results, two conclusions can be made:

1) In order to maximize the % correct score, forecast rain only when the probability is 50% or more. For this data the optimum threshold was actually 60%. Hughes and Sangster (5) attribute this to the fact that the 50% probability forecasts were unreliable since the observed relative frequency of rain (when a 50% probability was forecasted) was something other than 50%. Normally one can expect the optimum threshold to be 50%.

2) In order to maximize the threat score, the threshold should be lower. For this data one should forecast rain only when the probability was 30% or more.

From this data, we see that regardless of which of the two scores the forecaster is trying to maximize, rain should be

forecasted when the probability is 50% or more, and no rain should be forecasted when the probability is lower than 30%. The real problem arises when the forecasted probability is between 30% and 50%. Should a forecaster predict rain in this situation? The answer depends on which score the public is using to evaluate their forecasters. If all the public desires from the forecaster is that he be right as often as possible, then no rain should be forecast in this situation. However, if the public wants to be warned about possible rain events (i.e., if the public considers rain events to be more important than non-rain events), then the forecaster should predict rain in this instance. If the public does not know which score the forecaster is trying to optimize, then misinterpretation of the forecast is likely.

5. DISCUSSION

Just as it is important to be able to optimize a skill score once probability estimates have been made, it is equally important to be able to correctly interpret an objective model forecast when making the initial probability estimates. Objective forecast models such as Model Output Statistics (MOS) (6) are designed to do well on a particular skill score. It is up to the forecaster to understand which skill score the model is trying to optimize when making his initial probability estimates. If a forecast contest is being evaluated by a skill score that is different from the skill score an objective model is trying to optimize, then the forecaster must take this into account. Neglecting this information will lead to a poor performance in the contest. For example, in the National Collegiate Weather Forecast Contest (precipitation categories are given in section two), if the MOS categorical precipitation forecast model is predicting .25–.49″ of precipitation, many forecasters will say "look at this, MOS is forecasting category 3 precipitation!" While this is true in a literal sense, it shows that many forecasters are not taking into account the fact that MOS categorical precipitation forecasts are designed to optimize the threat score (7). Thus, MOS will have a wet bias with respect to the [forecast-observed] skill score that is used to evaluate forecasts in the National Collegiate Weather Forecast Contest. So when one converts the MOS forecast from the value which optimizes the threat score to a value which optimizes the threat score to a value which optimizes the [forecast-observed] score, the forecaster will readily understand that MOS is "really" forecasting category 2 precipitation (with respect to the way the National Collegiate Weather Forecast Contest gets evaluated).

6. Conclusions

In summary, successful forecasting of surface weather elements begins with the use of meteorological information to generate reliable probability forecasts. Statistical models such as MOS are known to generate reliable probability forecasts for many weather elements. After probability estimates have been made, the forecaster should use them to select a forecast that is consistent with the directives of the skill score by which the forecast will be evaluated. The directives of four skill scores were given above and it was shown that forecast skill can diminish quite rapidly with departure from the appropriate directive.

ACKNOWLEDGMENTS

We thank J. M. Fritsch for reviewing this paper.

NOTES AND REFERENCES

1. Robert L. Vislocky received his B.S. in 1985 at the Department of Meteorology and Physical Oceanography of Rutgers University, and his M.S. in 1988 at the Department of Meteorology of the Pennsylvania State University. He is currently a doctoral candidate at the Department of Meteorology of the Pennsylvania State University. Recent interests include the statistical interpretation of numerical model output for weather forecasting.

2. George S. Young received his B.S. in 1979 and his M.S. in 1982 at the Department of Meteorology of Florida State University, and his Ph.D. in 1986 at the Department of Atmospheric Sciences of Colorado State University. He is currently (1986-present) a member of the faculty of the Department of Meteorology of the Pennsylvania State University. Recent interests include the interaction of boundary layer turbulence with mesoscale circulations. Present address: Department of Meteorology, 503 Walker Building, The Pennsylvania State University, University Park, Pennsylvania 16802.

3. Murphy, A.H., and R.W. Katz, 1985: Probability, Statistics, and Decision Making in the Atmospheric Sciences. *Westview Press, Boulder, Co., 545 pp.*

4. Sanders, F., 1967: The verfication of probability forecasts. Jrnl. of Appld. Meteor., *Vol. 6, 756–761.*

5. Hughes, L. A., and W. E. Sangster, 1970: A note on the categorical verification of probability forecasts. ESSA Tech. Memo. WBTM CR-35, *Kansas City, Mo., 10 pp.*

6. Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. Jrnl. of Appld. Meteor., *Vol. 11, 1203–1211.*

7. Bermowitz, R. J., and E. A. Zurndorfer, 1979: Automated guidance for predicting quantitative precipitation. Mon. Wea. Rev., *Vol. 107, 122–128.*