NOTES

A Bias in Skill in Forecasts Based on Analogues and Antilogues

H. M. VAN DEN DOOL

Department of Meteorology, University of Maryland, College Park, MD 20742

21 November 1986 and 28 January 1987

ABSTRACT

A bias in skill may exist in statistical forecast methods in which the verification datum is withheld from the developmental data (cross-validation methods). Under certain circumstances this bias in skill can become troublesome. By way of example, it is shown that the judgment of the quality of forecasts based on analogues and anti-analogues may severely suffer from a bias in skill. A cure to the problem is discussed. Some implications for published results of long-range weather forecasting models based on analogues are discussed.

1. Introduction

The problem of artificial skill in statistical forecasting is well known. A forecast variable Y may be related to the variable X in a give dataset, but on independent data the previously established empirical relationship does not always hold up. The shrinkage of skill is especially problematic if the relationship between Y and X must be discovered from the dependent data, or, if Y is forecast from (too) many predictors ($X_1, \ldots X_n$) in a multiple-regression equation. Recent discussions of the artificial skill problem can be found in Lanzante (1984), Michaelsen (1985), Shapiro and Chelton (1986) and Lanzante (1986), among others.

There are at least two reasons for artificial skill: (i) Sampling fluctuations are taken for real and (ii) there may be errors in the design of the forecasting experiment resulting in "skill" in the absence of any real skill. In this note we want to discuss a peculiar source of artificial skill that falls in category (ii), but which is caused by honest attempts to minimize sampling errors.

Many workers are aware of the dangers of sampling and how misleading an a posteriori forecast relationship derived from a limited sample can be. In order to avoid this trap, several strategies have been developed. The *first* is to keep some (usually recent) data in reserve as independent data in order to test a relation derived from historical data. Examples are Barnett (1981), Klein (1983), and Bhalme et al. (1986). However, since samples are often small (~35 yr for models involving upper air data), no one in the field of long-range forecasting has the luxury of plentiful dependent and independent data. Therefore a *second* method is becoming popular, in which each datum is used for both development and verification. The idea is to withhold the verification datum from the developmental data set and to cycle repeatedly through this process while changing the verification datum (and the developmental dataset) slightly. The second method obviously consumes a lot of computer time. Withholding the verification data has been named "cross-validation," and examples are Michaelsen (1985) and Dixon and Harnack (1986).

We will now show, by way of example, how artificial skill can become a problem if the withholding of the verification datum is not properly done. The example is taken from the field of long-range weather prediction (LRWP), where skill scores are notoriously low and artificial skill is more often a serious problem that at lead times where true skills are much larger. Given 55 yr of monthly mean surface air temperature at 344 Climate Divisions in the United States, we will search for analogues, anti-analogues (the latter contracted to antilogues from now on) and discuss the skill of 1-month forecasts based on the subsequent weather in ana- (anti-) logue years. The analogue method is similar to cross validation in that it withholds the target month from the pool of potential analogues. (This is obvious, since the target year is an unbeatable perfect match to itself.)

In section 2 we discuss the data and the analogue selection and verification procedures. Examples of incorrect and correct withholding of the verification data are given in section 3. Finally, a summary and concluding remarks are presented in section 4.

2. Data, selection and verification

The data used in this study consist of 55 yr of monthly mean air temperatures (MMAT) at 344 United States Climate Divisions (CD) for the years 1931–85. (January 1986 is included for verification, as well.) This dataset has recently been published in map form by Cayan et al. (1986). Here we use the original data, including the potentially suspect CDs that were identified in Cayan et al.'s introduction.

The MMAT given at 344 CDs (s), 12 months (m) and 55 yr (j) will be denoted by $T(s, m, j)$. Taking out

the 55-yr mean for each month and dividing by the appropriate temporal standard deviation transforms the raw $T(s, m, j)$ into the standardized anomalies $\hat{T}(s, m, j)$ depicted in Cayan et al. One advantage of standardized anomalies is that it places areas of high and low variance on an equal footing.

Similarity of two $\hat{T}$ fields is expressed by a type of pattern correlation coefficient rather similar to that of Bergen and Harnack (1982), i.e.,

$$PC^*(m1, j1; m2, j2) = \frac{\frac{1}{N}\sum_s \hat{T}(s, m1, j1)\hat{T}(s, m2, j2)}{\left\{\frac{1}{N}\sum_s [\hat{T}(s, m1, j1)]^2 \frac{1}{N}\sum_s [\hat{T}(s, m2, j2)]^2\right\}^{1/2}}. \quad (1)$$

Equation (1) gives a measure of similarity of the two $\hat{T}$ fields over the entire contiguous United States, and gives equal weight to all CDs; $N = 344$ indicates the number of CDs. For a base year, denoted by jb, and month m, the best analogue ($ja1$) and best antilogue ($jt1$) are determined by $PC^*(m, jb; m, ja1) \geq PC^*(m, jb; m, j)$ for all $j \neq jb$, and $PC^*(m, jb; m, jt1) \leq PC^*(m, jb; m, j)$ for all $j \neq jb$; PC^* measures the quality of the ana- and antilogue. Similar to Barnett and Preisendorfer (1978), the ana- (anti-) logue is not constrained to be in the past. The second best analogue year $ja2$ is defined similarly from $PC^*(m, jb; m, ja2) \geq PC^*(m, jb, m, j)$ for all $j \neq jb$ and $j \neq ja1$ and so on for the third, fourth, etc., analogue and antilogue.

The forecast for base year jb's next month based on the best analogue is $\hat{T}(s, m + 1, ja1)$, based on the best antilogue the forecast is $-\hat{T}(s, m + 1, jt1)$ and so on. Forecasts are verified by the very same PC^*, that is, $PC^*(m + 1, jb; m + 1, ja1)$ for the best analogue, and so on.

Forecasts based only on the best analogue may be subject to large sampling error. Therefore, forecasts derived from a combination of good analogues have to be considered (Bergen and Harnack, 1982). The way in which N analogues are combined here is given by

$$\hat{T}(s, m+1) = [PC^*(m, jb; m, ja1)]^2 \hat{T}(s, m+1, ja1) + \cdots [PC^*(m, jb; m, jaN)]^2 \hat{T}(s, m+1, jaN)/W,$$

where W is the sum of the weights. The procedure is similar for a combination of antilogues, with a change of sign at the end of the summation. A mixture of analogues and antilogues can be similarly created by changing the sign of each of the antilogues and summing them in with analogues. Assuming some degree of linearity, this mix of N analogues and antilogues should give the best forecasts because the quality of the best N ana- (anti-) logues is higher than that of the N analogues or N antilogues alone.

3. Results

The analogue system previously defined has a small but demonstrable forecast skill. The results are best in summer and winter, but for our discussion it suffices to present yearly mean skill only. Figure 1a shows the yearly mean skill of monthly forecasts based on analogues (curve ANA), antilogues (curve ANTI) and a mixture of analogues and antilogues (curve MIX) as a function of the number (N) of ana- (anti-) logues used. The results are rather surprising. Only the mixed system acts according to expectation, that is, increasing skill with N and leveling off at large N (at 10% skill or so). It is shocking to see that (i) antilogues do so much better than analogues, (ii) the skill of antilogues keeps increasing with N even at large N and (iii) the skill of analogues returns to near zero at large N.

None of the items (i)–(iii) is related to an interesting physical phenomenon, but rather to a flaw in the design of the experiment. The flaw is that the 55-yr mean has been removed prior to the analogue searching and forecast verification. By definition

$$\sum_{j=1}^{55} \hat{T}(s, m, j) = 0$$

and therefore,

$$\hat{T}(s, m, jb) = -\sum_{j=1, j\neq jb}^{55} \hat{T}(s, m, j). \quad (2)$$

As a result, there is a tendency for negative correlations between $\hat{T}(s, m, jb)$ and $\hat{T}(s, m, j)$, $j \neq jb$, and thus a tendency for assigning higher quality to antilogues. By combining 54 antilogues, one could end up with a perfect forecast ($PC^* = 100\%$).[1] In summary, the problem is that antilogues are preferred in the selection, that combinations of antilogues will have a strong positive bias in skill, that combinations of analogues will have a strong negative bias in skill and that the mixed system has a moderate positive bias in skill.

Of course, the magnitude of the problem depends very much on the length of the dataset. Here we are lucky enough to have 55 yr. Even with 55 yr, however, the problem is quite bad. The magnitude of the problem also depends on the maximum number of antilogues allowed in the forecast (here $N = 15$), the way in which analogues and antilogues are combined [here $(PC^*)^2$], and the way the forecast is verified (here PC^*).

The cure is easy, but computationally intense. For searching analogues for January 1962, the standard deviation and the mean should be calculated over 1931–61 and 1963–85, leaving out 1962. The $T(s, 1,$

[1] Ranking from the largest negative correlation upward, there are 54 antilogues. Number 54, the worst antilogue, has the highest positive correlation and is the best analogue at the same time.

j), j = 1931–85 (including 1962) is then standardized using the 1931–61, 1963–85 mean and standard deviation. The verification data $\hat{T}(s, 2, 1962)$ should be treated similarly, excluding 1962 from the calculation of the mean and the standard deviation. When this is done, Fig. 1a is replaced by Fig. 1b. The results are now quite acceptable. All curves level off nicely at large N, the antilogue and analogue method have about the same skill, and a mix of analogues and antilogues has the best skill of the three "models" considered.

We conclude that due to an incorrect treatment of the target year, the pure analogue method has a large negative bias in skill, while the pure antilogue method has a similarly large positive bias in skill. Note also, by comparing Figs. 1a and 1b, that the mixed method has only a small positive bias, about 10% of the true skill.

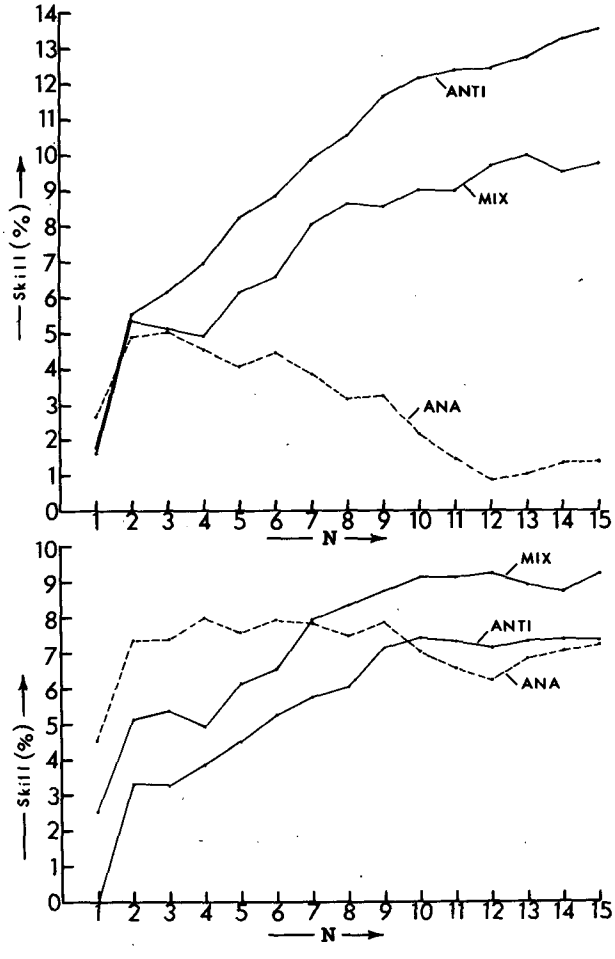

FIG. 1a. The skill of forecasts of monthly mean temperature anomalies over the United States based on analogue and anti-analogue methods as a function of the number analogues (N) combined to produce the forecast. The skill is measured by the pattern correlation coefficient (%), see Eq. (1). The curve labeled ANA is based on combining analogues only, the curve ANTI on antilogues only and the curve MIX is based on both ana- and antilogues.

FIG. 1b. As in Fig. 1a, but now the temperature of the target year is withheld in calculating the time mean. See text for more details.

4. Conclusion and discussion

The results presented in this note indicate that in forecast methods in which the verification datum is withheld from the development data, it is imperative to calculate the mean and the standard deviation from the developmental data only, and to treat the verification datum as totally independent.

In the published literature there are quite a few examples in the field of long-range weather forecasting where the data were standardized prior to the application of cross-validation forecast techniques. This causes the verification data to be negatively correlated with the developmental data. Whether or not this has led to a serious bias in the verification scores reported in these papers is hard to tell because the bias is a function of the sample size, the forecast method in question and the skill scoring method. In the present note we have only shown, by example, that in the ana- (anti-) logue method there is a bias in skill which may become very serious indeed, especially if one combines ana- (anti-) logues. These conditions must have been similarly conducive to producing a bias in the skill of the analogue forecast methods reported by Barnett and Preisendorfer (1978) and Bergen and Harnack (1982). Since only analogues were used in these two studies, their results must have had a negative bias. Unpublished work by Barnston and Livezey at the Climate Analysis Center confirms the suspicion that the Barnett and Preisendorfer analogue system does indeed have a nonnegligible bias in skill (using a tercile scoring method). And it may very well be true that, upon extending the Bergen and Harnack method, the conclusion presented by Harnack et al. (1985) that "it is clear in general that skill levels improve when negative analogs were used as input in addition to positive analogs. Other results indicate that using three additional positive analogs . . . does not likewise improve skill" can be attributed entirely to the inclusion of the verification datum in calculating the mean that was subtracted from all data prior to the analogue selection and verification.

No matter how small or large the problem is, it is always better to avoid all risks by recalculating means and standard deviations from the developmental data excluding the verification datum. This is quite time consuming, however. Special attention should be paid when data have to be detrended, as well. In that case, withholding data will create a gap in the time series to which a trend line has to be adapted.

None of the above applies to operational forecasts where models are always based on past data and the verification data, unknown to us, can not be erroneously included in the mean and the standard deviation.

Acknowledgments. This note is the result of stimulating discussions with Tony Barnston. The work was supported by the Cooperative Institute of Climate Studies under NOAA Grant NA84-AA-H-00026.

REFERENCES

Barnett, T. P., 1981: Statistical prediction of North American air temperatures from Pacific predictors. *Mon. Wea. Rev.*, **109,** 1021–1041.

——, and R. W. Preisendorfer, 1978: Multifield analog prediction of short-term climate fluctuations using a climate state vector. *J. Atmos. Sci.*, **35,** 1771–1787.

Bergen, R. E., and R. P. Harnack, 1982: Long-range temperature prediction using a simple analog approach. *Mon. Wea. Rev.*, **110,** 1083–1099.

Bhalme, H. N., S. K. Jadhav, D. A. Mooley and Bh. V. Ramana Murthy, 1986: Forecasting of a monsoon performance over India. *J. Climatol.*, **6,** 347–354.

Cayan, D. R., C. F. Ropelewski and T. R. Karl, 1986: An atlas of United States monthly and seasonal temperature anomalies December 1930–November 1984. NOAA—U.S. Climate Program Office, 244 pp.

Dixon, K. W., and R. P. Harnack, 1986: The effect of intraseasonal circulation variability on winter temperature forecast skill. *Mon. Wea. Rev.*, **114,** 208–214.

Harnack, R., M. Cammarata, K. Dixon, J. Lanzante and J. Harnack, 1985: Summary of U.S. seasonal temperature forecast experiments. *Proc. of the Ninth Conf. on Probability and Statistics in Atmospheric Sciences.* Virginia Beach, 175–179.

Klein, W. H., 1983: Objective specification of monthly mean surface temperature from mean 700-mb height in winter. *Mon. Wea. Rev.*, **111,** p. 674–691. Lanzante, J. R., 1984: Strategies for assessing skill and significance of screening regression models with emphasis on Monte Carlo techniques. *J. Climate Appl. Meteor.*, **23,** 1454–1458.

——, 1986: Reply. *J. Climate Appl. Meteor.*, **25,** 1485–86.

Michaelson, J., 1985; Estimation of artificial skill in forecast models. *Proc. of the Ninth Conf. on Probability and Statistics in Atmospheric Sciences,* Virginia Beach, 247–251.

Shapiro, L., and D. Chelton, 1986: Comments on "Strategies for assessing skill and significance of screening regression models with emphasis on Monte Carlo techniques." *J. Climate Appl. Meteor.*, **25,** 1295–1298.