A Degeneracy in Cross-Validated Skill in Regression-based Forecasts

ANTHONY G. BARNSTON AND HUUG M. VAN DEN DOOL

NWS/NMC/Climate Analysis Center, Washington, D.C.

(Manuscript received 10 June 1991, in final form 19 May 1992)

ABSTRACT

Highly negative skill scores may occur in regression-based experimental forecast trials in which the data being forecast are withheld in turn from a fixed sample, and the remaining data are used to develop regression relationships—that is, exhaustive cross-validation methods. A small negative bias in skill is amplified when forecasts are verified using the correlation between forecasts and actual data. The same outcome occurs when forecasts are amplitude-inflated in conversion to a categorical system and scored in a "number of hits" framework. The effect becomes severe when predictor–predictand relationships are weak, as is often the case in climate prediction. Some basic characteristics of this degeneracy are explored for regression-based cross-validation.

Simulations using both randomized and designed datasets indicate that the correlation skill score degeneracy becomes important when nearly all of the available sample is used to develop forecast equations for the remaining (very few) points, and when the predictability in the full dependent sample falls short of the conventional requirement for statistical significance for the sample size. The undesirable effects can be reduced with one of the following methodological adjustments: 1) excluding more than a very small portion of the sample from the development group for each cross-validation forecast trial or 2) redefining the "total available sample" within one cross-validation exercise. A more complete elimination of the effects is achieved by 1) downward adjusting the magnitude of negative correlation skills in proportion to forecast amplitude, 2) regarding negative correlation skills as zero, or 3) using a forecast verification measure other than correlation such as root-mean-square error.

When the correlation skill score degeneracy is acknowledged and treated appropriately, cross-validation remains an effective and valid technique for estimating predictive skill for independent data.

1. Introduction

The desire to accurately quantify statistical forecast skill has existed among meteorologists and oceanographers for many years. Estimates of predictive skill based on a posteriori data-fitting techniques such as regression using limited samples are characteristically higher than the skill would be in the population from which the sample is drawn. This is reflected in the generally lower skill levels obtained when the sample equations are used to forecast future or otherwise independent data.

In order to reduce the problem of artificial skill produced from overfitting and thus receive a more representative estimate of real skill, researchers have used *cross-validation* methods, in which forecast models (regression or other) are developed using only part of the available dataset and then applied to the independent data points left out. The number of points left out can range from one to more than half of the available dataset, and the set of removed points may be changed so that a large number of forecasts (perhaps for all possible combinations of a given number of withdrawn points) can be made. With the relatively recent introduction of larger computers, cross validation has come to imply such exhaustive or quasi-exhaustive trials. Discussion and/or examples of cross-validation in association with regressionlike approaches are found in Klein (1983), Van den Dool (1984), Harnack et al. (1985), Dixon and Harnack (1986), Michaelsen (1987), Barnett and Preisendorfer (1987), and Livezey et al. (1990) (the Klein and Van den Dool studies do not use the exhaustive version). Cross validation has also been used in analog forecasting (Barnett and Preisendorfer 1978; Livezey and Barnston 1988).

When cross-validation is applied under appropriate circumstances (i.e. using data that are stationary and not significantly autocorrelated), it generally produces scores representing the skill expected on application of the sample relationship to independent data—that is, lower skill scores than those produced in full dependent sample analyses. However, there are features in the design of cross-validation experiments that can introduce biases or degeneracies in the results under some circumstances, unless specific preventive measures are taken. An example of a flawed design leading to substantial artificial skill in certain types of analog forecasts

Corresponding author address: Dr. Anthony G. Barnston, Climate Analysis Center, W/NMC51, World Weather Building, Rm. 604, 5200 Auth Road, Camp Springs, MD 20746.

using cross-validation is discussed by Van den Dool (1987). In this study another example of a design flaw is presented.

This study identifies, illustrates, explains, and provides solutions to the problem of the perhaps unexpected appearance of highly negative skill scores using a regression-based cross-validation design when the skill score is computed as a correlation, or percent variance explained, between the set of forecasts and the corresponding set of actual data. This also occurs when regression-based forecasts are converted to a categorical forecast system (and scored using a "number of hits" skill measure such as the Heidke score), if the original amplitude of the forecasts is inflated. In either case, it occurs in low predictive skill environments in which a near-zero skill score would be reasonably expected. Awareness of the problem began with the occurrence of highly negative correlation-based skill scores in ongoing prediction research at the Climate Analysis Center. It is considered a problem both for aesthetic reasons (e.g., the appearance of intensely negative "bull's-eyes" on spatial maps of forecast skill) and for more substantive reasons (e.g., the computation of a mean skill score over a spatial domain).

In section 2 the cross-validation procedure is described. Section 3 provides illustrative examples of the degeneracy under study; subsections 3a and 3b are most vital here. In section 4 some features of the degeneracy are explained quantitatively, the difference between this phenomenon and "ordinary" artificial skill is discussed, and solutions to the degeneracy problem are presented. A summary and some conclusions are given in section 5.

2. Cross-validation procedure

In an exhaustive cross-validation experiment, a sample of data is first defined, consisting of one predictand and one (or more) predictors and a prescribed number of paired elements (predictor–predictand "points"—e.g., each point representing climate data for, say, a particular year). Typically the sample is drawn from a population with a potentially infinite number of points, and is the largest possible sample given the practical considerations of the real world. Then the cross-validation exercise is performed in which a given number of points are withheld from the sample, a regression equation is developed using only the remaining points, and forecast(s) is made for the value(s) of the predictand of the withheld point(s), using the predictor value(s) of the latter point(s). Before forming the regression equation, the developmental points are restandardized in terms of their own mean and standard deviation, and the withheld points should be standardized in terms of these same newly computed statistics for the verification process (Van den Dool 1987). The exercise is repeated exhaustively such that each possible combination of the given number of excluded points is used as the forecast target. Finally, a correlation (or other skill score such as root-mean-square error or mean absolute error) between forecasts and verifications is computed.

3. Examples

The degeneracy under examination only occurs systematically when a *fixed* "full sample" of data is defined for cross-validation testing. In actual or operational forecasting it is not observed, because the full sample continues to change as each new datum is acquired. In this section we examine the degeneracy in various simulated cross-validation conditions.

In performing the simulations, two types of datasets are used. The first type is random, in which case a random number generator is called to supply numbers from a Gaussian distribution to form the predictor (x) and predictand (y) groups of size N. In this case it is necessary to perform many iterations, each using a different pair of random datasets, because any individual pair does not contain exactly Gaussian x and y and could produce unrepresentative results. The performance of many iterations not only provides a more realistic mean result but also provides useful estimates of the distribution (i.e., variability) of the results.

The second type of dataset used in the experiments is the designed dataset, in which a single prototype or textbook version of a dataset is constructed. For example, a bivariate Gaussian distribution might consist of symmetric rings of points in the x, y plane positioned about the origin according to the Gaussian probability density. A disadvantage here is that no distributional information is provided in the test results. The advantage is that an approximate mean result can be obtained more quickly. The representativeness of a designed dataset is verified through comparison of its results with the mean of the results using a large number of randomized datasets. Once verified, the designed dataset is trusted in more computer-intensive experiments for which only the mean result is desired.

a. Prototypical illustration

The simplest cross-validating experiment illustrating the behavior under study uses a designed dataset with four pairs of predictor–predictand (x, y) values: $(1, 1)$, $(1, -1)$, $(-1, 1)$, and $(-1, -1)$. The correlation between x and y is zero. An unbiased cross-validation skill-estimating technique would be expected to produce a skill of zero in predicting y from x from this dataset if each of the four points were withdrawn and used as the forecast target in turn. However, a cross-validation correlation of -1.0 is the result, implying that the forecasts are always of opposite anomaly sign to reality with a perfectly predictable amplitude relationship. This result comes about not because of the smallness of the sample; in fact, using 100 points at

each of the four locations also produces a −1 correlation. This occurs because when any one point is withheld as the forecast target, the remaining points no longer have zero correlation but have a correlation with sign in keeping with an axis of scatterplot elongation perpendicular to that of the line connecting the origin to the removed point. In the four-pair experiment, for example, when the point (1, 1) is removed, a negative correlation (−0.50) is produced among the remaining three points. When the predictor value (x) of the removed point is introduced in the resulting regression equation, the forecast for the predictand value ($\hat{y}$) of the point is half the correct amplitude and has the incorrect sign. The amplitudes of the forecasts for the other three points are similarly half the true amplitude and have the incorrect sign, and a −1 correlation between forecasts and observations appears. When the sample size is progressively increased for the same four locations, the correlation among the remaining points when one case is withdrawn becomes increasingly smaller, the regression coefficients and the amplitude of the forecasts approach zero, but the signs of the equally weak forecasts are still opposite those of their observational counterparts; thus, the correlation between forecasts and observations again is −1. This effect is clearly outside the expected sample size–related statistical variability. In fact, the mean skill value itself is altered from that expected in a zero-skill forecast environment, due to an inherent feature of the cross-validation design.

The near-zero amplitude forecasts of the larger sample experiments would verify with approximately zero skill (i.e., skill near that of a climatology forecast) rather than highly negative skill if an rms error (RMSE) score or a categorical hit versus miss score (as in Dixon and Harnack 1986) were used rather than the forecast versus observation correlation measure (for categorical forecasts this near-zero skill expectation holds only in the absence of forecast amplitude inflation; i.e., no forecasts would fall in the outer categories).

The above example is contrived and simplified. However, its behavior occurs to varying degrees in real world regression-based cross-validation, creating a problem different from the more familiar artificial skill (shrinkage) phenomenon discussed in Davis (1976) and elsewhere. We next examine the seriousness of this effect (to be called a degeneracy) for realistic values of the major parameters of regression and cross-validation, given various "true" values of the full sample correlation between predictor(s) and predictand.

b. *Dependence on sample size*

The cross-validation simulation of the type described in subsection 3a can be extended to different initial x versus y full sample correlations by moving each point closer to or farther away from the 45° $y = x$ line along its perpendicular to that line by a fixed proportion of its initial perpendicular distance from the line. The cross-validation skill can then be plotted as a function of the full sample correlation. Such a plot is not shown for the simple four-point designed (x, y) dataset discussed in subsection 3a, that dataset being used only to demonstrate an extreme case of the basic mechanism underlying the degeneracy.

To explore the effect of the number of full sample points N on this behavior, we use a set of samples of the more realistic normal distribution as drawn from a random number generator, and vary the sample size. In this experiment 200 iterations of 64 complete rounds of cross-validation (one round consisting of the holding out of each of the N points in the sample by itself once as the forecast target) are carried out, each iteration using an independent random data sample. Within each iteration the full sample correlation is systematically modified between −0.999 and 0.999 using the technique described above, resulting in 64 values having smaller differences (0.01) near zero, and larger differences (0.05) near −1 and 1.

Figure 1 shows the resulting relationship between the full sample (with no withheld points) correlation and the cross-validation correlation (reflecting predictive skill). This is shown for N = 16 (panel a), 32 (panel b), and 128 (panel c). In each of the three parts of Fig. 1 there are two groups of five roughly parallel curves, and also an uncurved $y = |x|$ line with vertex at the origin to highlight the expected relationship if no bias (including the skill score degeneracy) existed. The group of five curves that approximate the $y = |x|$ line for high full sample correlation magnitude and have deep, well-defined minima near $x = 0$ describe the cross-validation results using the correlation coefficient as the scoring measure. Within this group the solid curve represents the mean over the 200 iterations, the short dashed curves plus and minus one standard deviation from the mean, and the longer dashed curves the maximum and minimum results. (Note that the extreme curves can have noticeable "wiggles" because they may represent the outcome of one particular round of cross-validation for a given full sample correlation interval, and a different round for an adjacent interval.) The minimum value of the mean cross-validation skill occurs in the neighborhood of zero full sample correlation; this is also where the greatest discrepancy between the no-bias skill expectation and the actual skill occurs. This minimum value becomes somewhat stronger for smaller N. The degree of full sample correlation required to achieve a given smallness of the bias of the mean results decreases with sample size. This is in qualitative agreement with the inverse relationship of artificial hindcasting skill (due to overfitting) to sample size for a fixed number of predictor variables (Davis 1976; Michaelsen 1987). The dispersion of skill (reflected in the two sets of dashed lines) is substantial and is also inversely related to N. The full sample correlation required to obtain a non-

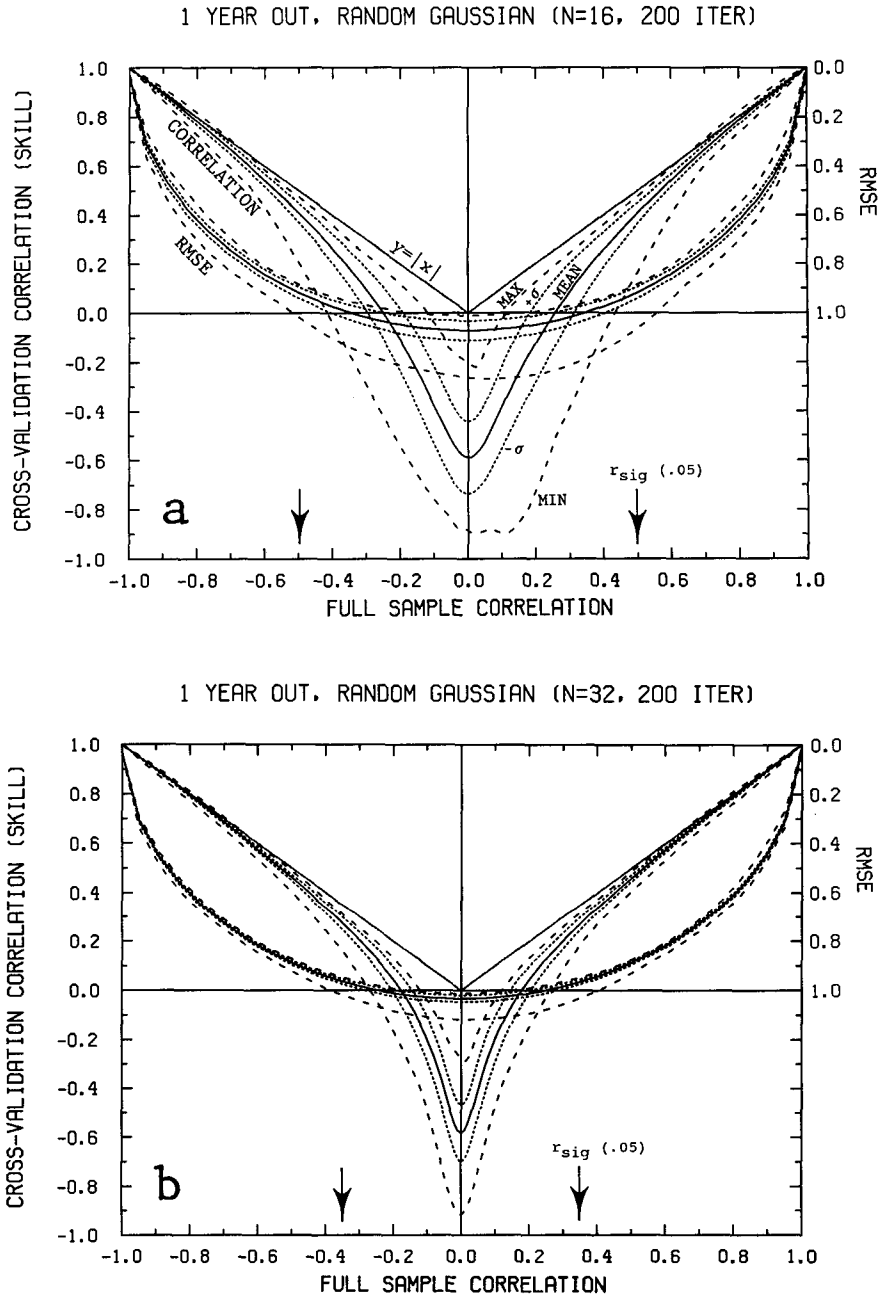

FIG. 1. Cross-validation correlation (correlation between forecasts and the data being forecast) as a function of the full sample correlation between predictor (x) and corresponding predictand (y) data for 200 successively drawn samples of x and y datasets from a random Gaussian number generator. The $y = |x|$ line shows expected skill in the absence of biases. The two groups of five approximately parallel curves in each of the three panels of the figure show experimental results where each point (or "year") is held out as the independent forecast target using full sample sizes (N) of 16 (panel a), 32 (panel b), and 128 (panel c). Within each panel, the group of five curves having more clearly defined minima near $x = 0$ show results using the correlation skill measure (left ordinate scale) and the five with shallower minima using RMSE (right ordinate scale). The full sample correlation value required for 0.05 significance is indicated on the lower abscissa. See text (sections 3b and 4a) for further detail.

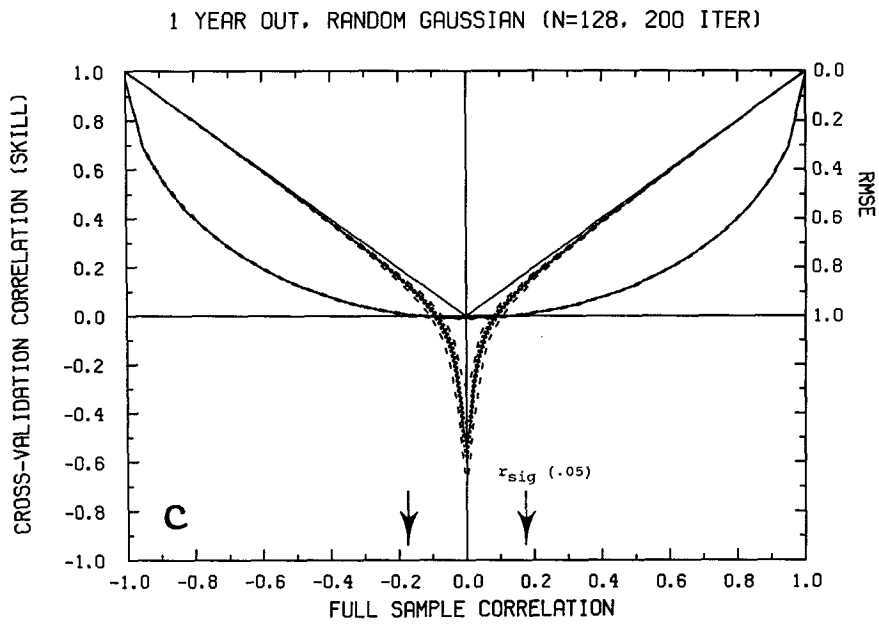

FIG. 1. (*Continued*)

negative mean cross-validation correlation skill appears to be roughly $\pm N^{-1/2}$ in these particular experiments. A more rigorous treatment of this characteristic of the degeneracy will be presented in section 4a.

The group of five curves having broader but shallower minima similarly represent cross-validation skill, but using RMSE rather than a correlation coefficient as the scoring measure (note ordinate scale on right side of plot, in which 1.0 represents the RMSE score of zero-anomaly or climatology forecasts). The $y = |x|$ line has no particular meaning with respect to the RMSE curves. Noteworthy features of the RMSE cross-validation results are 1) the comparatively small negative bias at near-zero full sample correlation associated with the low forecast amplitudes, and 2) the smaller variability of scores over the 200 random iterations than is found for the correlation skills, despite evidence of a negative skew for low full sample correlation values.

With the distributional features of the correlation-measured cross-validation skill well described in Fig. 1 for three sample sizes, further experiments are next performed to examine behavior under other specific conditions of the data, the cross-validation design, or the regression. In doing this, we are primarily interested in the mean skill results and thus no longer require evaluation of skill dispersion, which has been found to be fairly large for the correlation skill measure. For these additional tests, it is sufficient to use a single designed Gaussian dataset rather than a long series of random Gaussian datasets. For this purpose a 32-point designed bivariate Gaussian dataset is constructed (Table 1) and cross-validation correlation skill results are calibrated against the mean results shown in Fig. 1. The designed dataset result is found to be representative of the mean of the random dataset results, rendering it suitable for diagnosing the mean behavior in good approximation in further testing. In the process of performing these calibration tests it is found that minor differences in the designed dataset do not significantly change its simulated cross-validation skill results. Although the full sample correlation among the 32 (x, y) pairs in Table 1 is zero, it is modified systematically using the technique described at the beginning of this subsection.

TABLE 1. The individual points of the designed, approximately Gaussian bivariate dataset used for some of the cross-validation simulation experiments.

Point number	x	y	Point number	x	y
1	−0.2	−0.2	17	−0.8	−0.8
2	−0.3	0.0	18	−1.2	0.0
3	−0.2	0.2	19	−0.8	0.8
4	0.0	0.3	20	0.0	1.2
5	0.2	0.2	21	0.8	0.8
6	0.3	0.0	22	1.2	0.0
7	0.2	−0.2	23	0.8	−0.8
8	0.0	−0.3	24	0.0	−1.2
9	−0.5	−0.5	25	−1.3	−1.3
10	−0.7	0.0	26	−1.8	0.0
11	−0.5	0.5	27	−1.3	1.3
12	0.0	0.7	28	0.0	1.8
13	0.5	0.5	29	1.3	1.3
14	0.7	0.0	30	1.8	0.0
15	0.5	−0.5	31	1.3	−1.3
16	0.0	−0.7	32	0.0	−1.8

c. *Effect of outliers*

While the finer details of a dataset design may not affect the cross-validation skill behavior very much, outlier points (points falling relatively far outside the general distributional cluster) do significantly affect skill behavior (Michaelsen 1987). Examination of individual random data samples in the simulations whose results are shown in Fig. 1 reveals that the atypical cross-validation skills (e.g., those helping to form the minimum or maximum curves of the five-curve envelope) have marked asymmetry in either or both x and y, and sometimes one or more borderline outliers. Sometimes the full-sample correlation producing the minimum cross-validation correlation is somewhat removed from zero in those cases.

In three consecutive simulations using the 32-point designed Gaussian dataset discussed in subsection 3b (Table 1), the point (#29) initially located at (1.3, 1.3) is moved to (3.5, 3.5), to (5, 5), and finally to (10, 10). The 3.5 and 5 standard deviation anomalies are representative of climate events such as the 1982–83 El Niño or the Great Plains summer surface air temperatures in the hottest dust bowl summer, whereas the 10 standard deviation anomaly is performed only as an extreme design experiment. Note that upon restandardization for the cross-validation simulations with each point held out in turn, the standardized anomaly value of the outlier is substantially reduced—unless it is the point being withheld.

The result of the three outlier cross-validation simulations is shown in Fig. 2 along with that without an outlier. Compared with the outlier-free reference case, the outlier degenerative effects are more severe, and include 1) a general broadening of the full sample correlation interval within which the cross-validation correlation skill is negative, 2) a stronger negative minimum cross-validation skill, and 3) a migration away from zero (toward positive values in this example) of the full sample correlation producing the minimum skill. This is the full sample correlation value for which the outlier is most severely mispredicted when it is withheld (the other points contributing to a relationship of the opposite sign), and also for which the other points are most consistently mispredicted due to the outlier's strong opposing influence in the development subsample equation. The outlier creates a variation of the simple and extreme case of the complementary skill-destroying mechanism found in the four-point designed dataset discussed in subsection 3a. An outlier has an effect similar to that of decreasing the sample size by reducing the influences of and differences between the other points. This was evidenced in an ENSO cross-validation prediction study where a sea surface temperature persistence forecast was used as a skill control measure, and the 1982–83 El Niño produced an outlier in sea surface temperature persistence (Barnston and Ropelewski 1992).

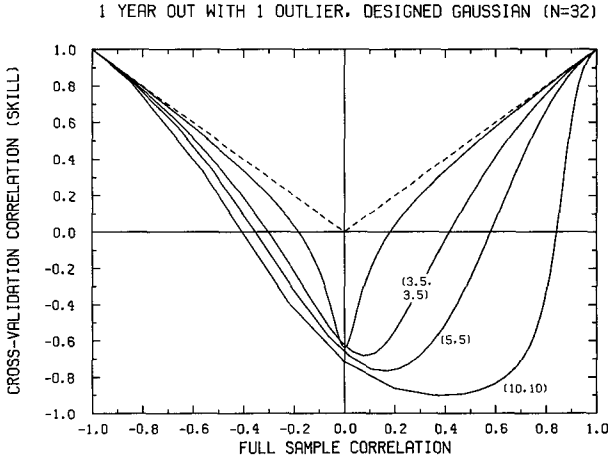

FIG. 2. Cross-validation correlation as a function of full sample correlation with each of three designed outliers replacing the point (1.3, 1.3) in a designed 32-point Gaussian dataset. The weakest outlier is placed in scatterplot location (3.5, 3.5), followed by two progressively more extreme outlier prescriptions: (5, 5) and (10, 10).

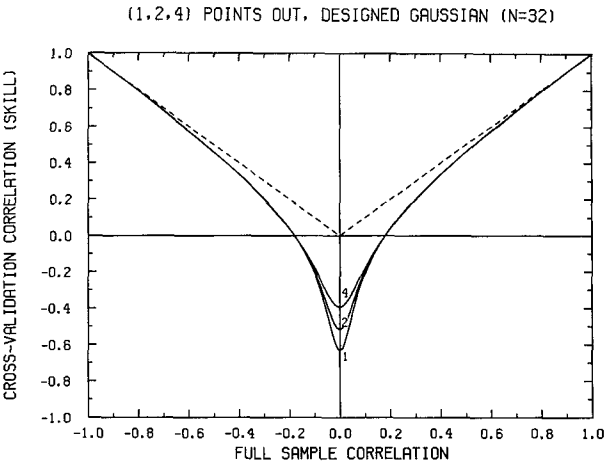

FIG. 3. Cross-validation correlation as a function of full sample correlation with a 32-point designed Gaussian density withholding one (bottom solid curve), two (middle), and four (solid curve nearest the dashed line) points (or "years") in the cross-validation procedure.

d. *Dependence on the number of points held out*

Figure 3 shows results using the designed 32-point Gaussian dataset for simulations holding out all possible combinations of one, two, and four points. For a full sample correlation of zero, the four-point removal results in the least severely negative cross-validation correlation (−0.41), followed by the two-point removal condition (−0.53) and the most severe one-point removal condition (−0.64). However, removing several more points at a time does not visibly affect the nonnegative cross-validation scores or the zero-skill crossing points, suggesting that only the behavior within the

negative skill score interval is substantially affected by the degeneracy-producing design flaw. The withholding of more than four points was carried out using a designed 12-point Gaussian dataset (to reduce the large computational burden). Results withholding 1, 2, 4, and 8 points out of 12 (not shown) reveal that removing even two-thirds of the sample further diminishes the degeneracy only within most of the range of the already negative correlation skill scores. For positive cross-validation skill scores the holding out of more points degrades skill and thus widens the interval of negative skill (particularly noticeable in the case of holding out four or eight points). This occurs as a result of the now noticeably proportionally smaller development subsample size, the effect of which was noted in Fig. 1 for the fixed condition of one point being withheld. (Although this occurs in the Fig. 3 experiments, it is not detectable because of the proportionally similar development subsample sizes.) These observations suggest a distinction between 1) ordinary, expected skill deflation when forecasting outside a development subsample of size N (Davis 1976; Michaelsen 1987) and 2) a special cross-validation methodologically based degeneracy primarily within the negative skill region.

e. Extension to multiple regression

Simulations are carried out for one point being withheld, using two and then five predictors in multiple regression to provide a ground for comparison to results using simple regression.

In performing the multiple regression simulations, the full sample correlation modifying technique (subsection 3b) is not used because of the complications surrounding the plurality of the predictors—that is, their own intercorrelations and the correlations of each with the predictand. Instead, the full sample multiple correlation is permitted to vary naturally by using many samples of three- or six-dimensional random Gaussian 32-point datasets in succession, each set having its own chance-produced predictors versus predictand multiple correlation. A disadvantage of this approach is the unavailability of high correlation magnitudes; the main concern, however, is with results at lower correlation magnitudes where the expected problem is found.

Results of the simulations are shown in Fig. 4, where panel (a) uses simple regression, (b) uses two-predictor regression, and (c) five-predictor regression. Because the multiple predictor coefficients may vary in sign, the square root of the multiple R squared value is used to describe the full sample correlation represented by the x axis. The strongly negative skill score degeneracy appears to increase with the number of predictors, as does the negative departure from the bias-free dashed line for the positive correlation skill scores. The latter effect is in keeping with the expectation of greater artificial hindcast skill for greater numbers of predictor variables due to increased opportunities for noise fitting (Davis 1976; Michaelsen 1987). The addition of predictors not only appears to worsen the degeneracy but also to increase its uncertainty as noted by the marked vertical scatter of points in Fig. 4c. Results of the Fig. 4b two-predictor regression simulation repeated for 64 points instead of 32 (not shown) reveal a somewhat lower average ordinate distance between the points and the dashed no-bias line for both positive and negative skills. An increase in the dataset sample size thus reduces both the degeneracy and, as expected, the amount of "normal" artificial skill (Davis 1976; Michaelsen 1987) in multiple as in simple regression-based cross-validation.

The multiple correlation cross-validation scoring degeneracy has been encountered in an ongoing long-range prediction study using the previous month's observed temperature and precipitation to parameterize mean soil moisture in predicting the present month's mean surface air temperature (Huang and Van den Dool 1993). For example, using 57 years of monthly mean data, Fig. 5 illustrates the exhaustive one-year-out cross-validation skill score degeneracy in the form of several strong negative cross-validation correlation skill pockets with a minimum value of −0.51 in North Dakota and Washington states for two-predictor multiple regression.

4. Discussion, explanation, and solutions

The examples discussed in section 3, while varied, have one feature in common: A full sample is defined, and all cross-validation trials use *all* data points in the full sample—either as part of the development subsample or part of the verification subsample. Exhaustive cross-validation involves great redundancy in the participation of each point—for example, in a full sample of size N where one point is held out in each forecast trial, each point is included in a development subsample N-1 times and in a verification subsample one time. These exhaustive, reciprocal features allow for a balanced and complete evaluation of the skill score degeneracy, and hopefully a realistic estimate of predictive skill in the theoretical infinite population of points.

a. The regression-based cross-validation skill score degeneracy

The rules of generalization of a regression relationship from a developmental subsample to a complementary verification subsample within a fixed full sample are quite different from the analogous rules from a sample to a population. The key difference is that in the latter case the population is inexhaustible, such that the selection and removal of any defined sample does not change the remaining population statistics or correlations.

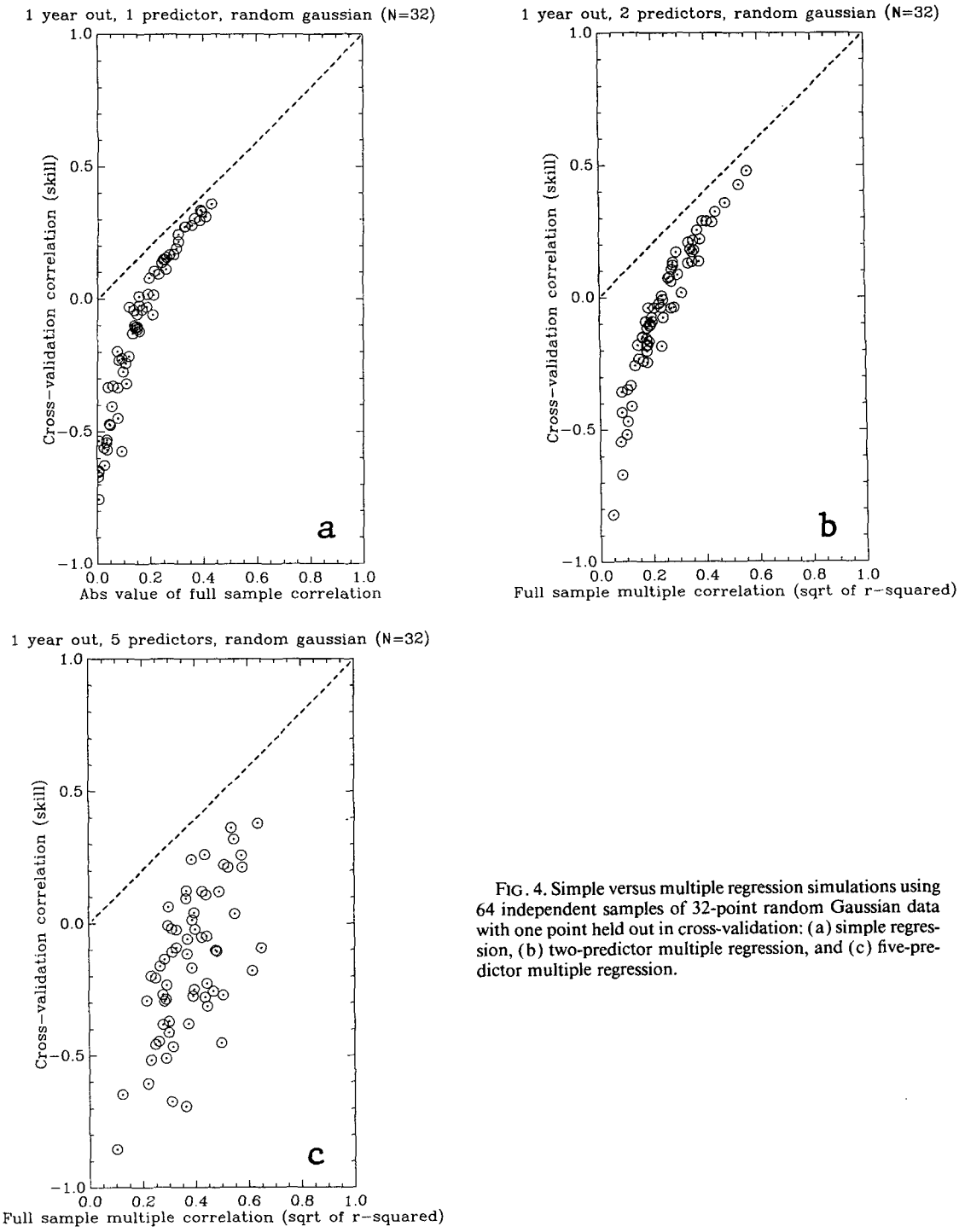

FIG. 4. Simple versus multiple regression simulations using 64 independent samples of 32-point random Gaussian data with one point held out in cross-validation: (a) simple regression, (b) two-predictor multiple regression, and (c) five-predictor multiple regression.

1) COVARIANCE INTERDEPENDENCE

When the full sample is used as a fixed "population" from which to select rotating sets of developmental elements and withheld elements for "independent" target tests, a complementary relationship arises between the predictor (x) versus predictand (y) statistics for points (x, y) in the development subsample as compared with those for point(s) (x, y) in the test sample. Suppose that one point is withheld for each

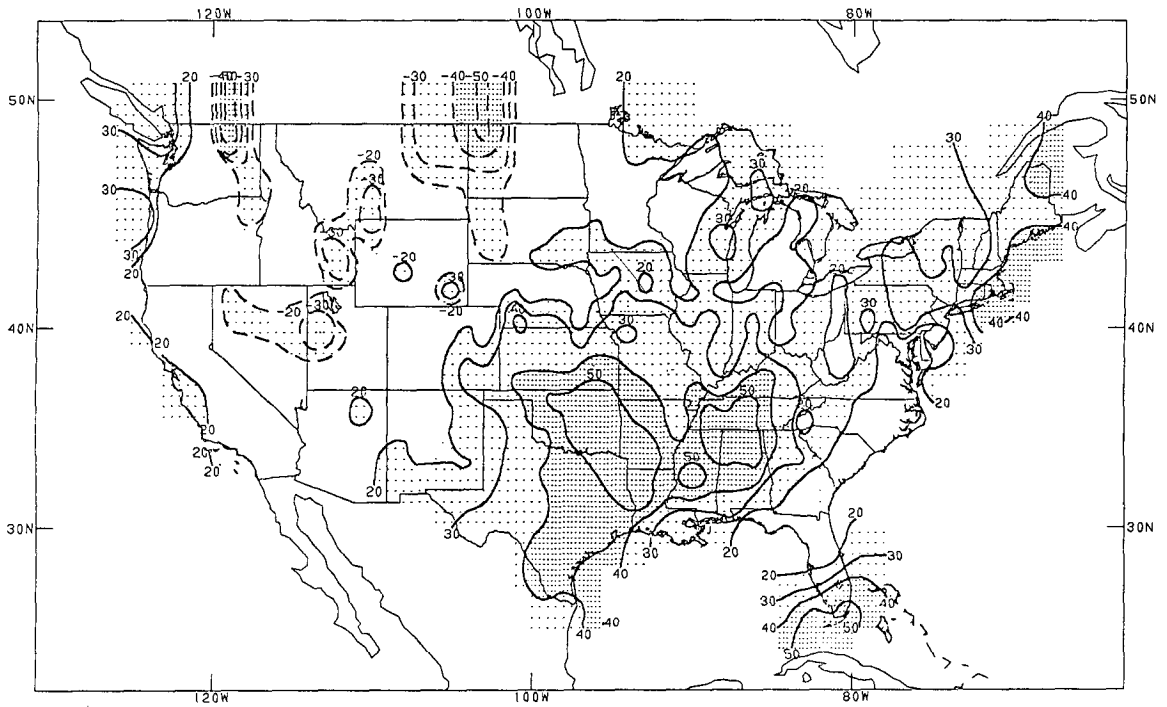

FIG. 5. Illustration of the subject degeneracy in a study using correlation-verified cross-validation, two-predictor multiple regression in prediction of monthly mean surface air temperature anomalies from previous month's temperature and precipitation anomalies (Huang and Van den Dool 1993). Here the geographic distribution of cross-validation skill is shown in predicting August temperature using July predictors. Each trial of cross-validation holds out 1 year as the forecast target and uses all remaining years to develop a regression equation. Units are correlation ×100. The −0.1, 0.0, and 0.1 contours are not shown. Areas of correlation skill score degeneracy are found in the northwestern United States with a minimum value of −0.51 in northwestern North Dakota and northeastern Washington.

forecast trial, and that the x and y are standardized. Let N denote the number of points in the full sample (fs) and j be the point number held out as the forecast target, leaving $N-1$ points in the development sample (ds). The full-sample correlation (r_{fs}) can then be expressed as

$$r_{fs} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i = \frac{1}{N}\Big(\sum_{\substack{i=1,N\\ i\neq j}} x_i y_i + x_j y_j\Big).$$

Let us assume for now that the restandardization of the points on the basis of the development sample statistics causes only minor changes in the cross-validation skill results (as will be discussed below, this is a realistic assumption except in very small samples or when an outlier is withheld). Thus, we approximate:

$$r_{fs} = \frac{1}{N}[(N-1)r_{ds} + x_j y_j]$$

or

$$\frac{1}{N}(x_j y_j) = r_{fs} - r_{ds}(N-1)/N. \qquad (1)$$

To obtain a contribution toward a positive cross-validation correlation skill in a single forecast trial within a cross-validation exercise, the left side of (1) must be of the same sign as r_{ds}; that is, the sense of the x versus y relationship in the withheld point must match that found in the development sample. When $r_{fs} = 0$, however, this can never happen, since, under these conditions, (1) reduces to

$$\frac{1}{N}(x_j y_j) = -r_{ds}(N-1)/N. \qquad (2)$$

Because the x_i and y_i generally assume a variety of nonzero values, exhaustive cross-validation using a fixed full sample is expected to result in negative correlation skill scores whenever $|r_{fs}|$ is less than some critical value r_{crit}.

The covariance interdependence here is analogous to the interdependence of means discussed in Van den Dool (1987) in the context of composite analog forecasting. In both cases the skill score is changed because of a design defect rather than for reasons of substantive interest. In the present case the design degeneracy is deceptive because it is easily mistaken for the artificial

skill effect related to sample noise fitting, as discussed in Davis (1976) and Michaelsen (1987).

2) "WIDTH" OF THE DEGENERACY

To determine how large an r_{fs} interval to expect to result in negative cross-validation correlation skill scores, we evaluate r_{crit} in terms of N (assumed mutually independent points) and certain important characteristics of the (x, y) distribution. From (1), we have

$$r_{ds} = \frac{Nr_{fs} - x_j y_j}{N - 1}. \tag{3}$$

The forecast $\hat{y}_j$ for any withheld y_j is given by $r_{ds}x_j$. Therefore,

$$\hat{y}_j = \frac{x_j(Nr_{fs} - x_j y_j)}{N - 1}. \tag{4}$$

If (4) is multiplied by the observed y_j and summed over all j, the result is

$$\sum_{j=1}^{N} y_j \hat{y}_j = \frac{Nr_{fs}}{N-1} \sum_{j=1}^{N} x_j y_j - \frac{\sum_{j=1}^{N} x_j^2 y_j^2}{N-1}.$$

The left-hand side is the sum of the cross products of forecasts and observations used to compute the cross-validation correlation skill score, and hence, carries the sign of that correlation. Using the fact that $\sum_{j=1}^{N} x_j y_j = Nr_{fs}$ for standardized x_j and y_j, we obtain

$$\sum_{j=1}^{N} y_j \hat{y}_j = \frac{N^2 r_{fs}}{N-1} r_{fs} - \frac{\sum_{j=1}^{N} x_j^2 y_j^2}{N-1}. \tag{5}$$

The requirement for the left-hand side to be positive is

$$\frac{N^2 r_{fs}^2}{N-1} > \frac{\sum_{j=1}^{N} x_j^2 y_j^2}{N-1},$$

or

$$|r_{fs}| > \left(\sum_{j=1}^{N} x_j^2 y_j^2/(N-1)\right)^{1/2} \frac{(N-1)^{1/2}}{N}. \tag{6}$$

Since the first factor on the right-hand side of (6), which we denote as k, is essentially independent of N, we find, approximately, $|r_{fs}| > kN^{-1/2}$. The proportionality factor k, for perfectly correlated x and y (or, equivalently, for x^4 or y^4), represents the square root of the coefficient of the fourth moment (kurtosis). For a standard Gaussian distribution this coefficient is equal to 3. However, for imperfectly correlated standard normal x and y the coefficient is lower, and drops to approximately unity for poorly correlated variables. It is also sensitive to deviations from normality in the x or y distributions—particularly to the tendency for occurrence of simultaneous outliers in x_j and y_j. To illustrate the degree of variability of k, its mean value over 200 cross-validation simulations, each using an independent set of 32 random normal (x, y) pairs, is 0.99 for $r_{fs} = r_{crit}$, with a standard deviation of 0.33 and extreme minimum and maximum occurrences of 0.43 and 2.29, respectively (these numbers all inflate slightly when x_j and y_j are restandardized using development sample statistics). This introduces some uncertainty to the determination of r_{crit}, since rarely in practice are the x and y distributions exactly Gaussian. Taking the square root helps reduce this uncertainty, which increases with decreasing N. To first order, then, we conclude that $|r_{crit}| = N^{-1/2}$. The r_{fs} intervals associated with negative cross-validation skill in the examples in section 3 approximately adhere to this rule. Near the center of the interval, skill scores become highly negative because of the interdependent complementary (proportionately opposite) covariance relationship noted in (2).

3) EQUATION FOR CROSS-VALIDATION CORRELATION SKILL

An approximate equation for the cross-validation correlation skill, r_{cv}, as a function of r_{fs} is developed by dividing (5) by N and by the standard deviation of the $\hat{y}_j$, or $\overline{|r_{ds}|}$ (the mean of $|r_{ds}|$ over N forecasts) so that the left-hand side equals r_{cv}. Then,

$$r_{cv} = \frac{\sum_{j=1}^{N} y_j \hat{y}_j}{N\overline{|r_{ds}|}} = \frac{Nr_{fs}^2}{\overline{|r_{ds}|}(N-1)} - \frac{k}{\overline{|r_{ds}|}(N-1)}.$$

For $|r_{fs}| > r_{crit}$, $\overline{|r_{ds}|}$ reduces to r_{fs} to close approximation. Assuming k roughly equals 1, the relationship for intermediate values of r_{fs} can be approximated:

$$r_{cv} = \frac{Nr_{fs}}{N-1} - \frac{1}{r_{fs}(N-1)}. \tag{7}$$

Except for where $|r_{fs}| < r_{crit}$, the curves of r_{cv} as a function of r_{fs} accompanying the examples in section 3 (e.g., Fig. 1) follow (7) reasonably closely. It should be noted that (7) describes the combination of degenerate behavior associated with covariance interdependence and the normal shrinkage associated with the use of fitted coefficients on data outside the development sample. As will be demonstrated in subsection 4b, the degenerate behavior plays a negligible role except in and near the $|r_{fs}| < r_{crit}$ interval.

The development of the relationships in this section without accounting for restandardization of the withheld points x_j and y_j turns out to generalize to the case of restandardization also. The difference in the x_j and y_j between the cases of not restandardizing and restandardizing is small except for very small or irregular

samples. In all cases except for zero initial anomaly, the anomaly magnitude of a restandardized x_j or y_j increases. The factor of increase decreases with N and increases nonlinearly as a function of the initial anomaly value. It is the differential inflation factor as a function of the initial anomaly value that affects cross-validation skill. The effect of restandardization for one point withheld is to slightly *reduce* the amplitude of the negative skill scores when r_{fs} is close to zero—by less than 1% for $N > 32$, about 1% for $N = 32$, and as much as 5% for $N = 8$. Away from the point of maximum degeneracy the effect of restandardization rapidly diminishes. For more than one point withheld the effect is slightly larger (e.g., 4% for $r_{fs} = 0$ for $N = 32$ for four points withheld) but under no condition approaches the magnitudes found for the composite form of analog forecasting discussed in Van den Dool (1987).

4) RELATIONSHIP TO STATISTICAL SIGNIFICANCE

As noted in section 4a(2), degenerate cross-validation results occur approximately within the interval $|r_{fs}| < r_{crit}$, where $r_{crit} = N^{-1/2}$. Because the full sample correlation required for statistical significance (r_{sig}) at any given significance level is also a function of $\sqrt{N}$, a correspondence between r_{sig} and r_{crit} is definable. Assuming that the sample is not very small, r_{sig} for the 0.05 level (two-sided) is about double r_{crit} for a priori selected predictor(s). The r_{sig} values for $N = 16$, 32, and 128 are indicated by arrows in the lower abscissa of each panel of Fig. 1. The statistical significance of r_{crit} itself is about 0.3 which is nonsignificant even by lenient standards. Because full dependent sample correlations achieving significance levels of 0.05 or stronger typically are not noticeably affected by the degeneracy, cross-validation in such cases can be trusted to provide representative estimates of the lower correlation skill expected when the development sample regression coefficients are applied to independent data.

b. *Cross validation without covariance interdependence*

The cross-validation skill result expected as a function of r_{fs} without the complementary covariance problem is found by conducting a simulation in which the interdependence between full sample correlation and that in the withheld point(s) is removed. This is approximately accomplished when a large (e.g., 7000 member) "population" of random bivariate normal predictor–predictand points is defined, and a cross-validation exercise is performed somewhat differently. Instead of the redundant, reciprocal procedure used in section 3, a simple regression relationship is derived using the first N points that is used to make a forecast for the next K points (see section 1 for the broad definition of cross-validation). An equation is then developed from the next N points (e.g., points $N + K + 1$ through $N + K + N$) and applied to the next K points, etc., until a sufficient number of iterations have been carried out to produce stable results. In this experiment the procedure is repeated 200 times using development sample sizes (N) of 31, 15, and 7 (comparable to full sample N of 32, 16, and 8 in the degeneracy experiments in section 3) and verification sample sizes (K) of 4. In each of the three cases, 800 forecasts are later verified using the correlation coefficient as in the examples in section 3. Note that the "full sample correlation" among each set of $N + 4$ points no longer remains constant from one set of forecast trials to the next. When the entire procedure is repeated for different values of the initial overall predictor–predictand correlation among the 7000 points, the resulting cross-validation correlation skill versus the "population" correlation (the abscissa of a plot, analogous to the full sample correlation in section 3) is as shown in Fig. 6. A much shallower cross-validation skill minimum in the near-zero "population" correlation region is noted in comparison to that of the degenerate simulations shown in section 3. In fact, when the population correlation is zero, there is no particular bias toward negative cross-validation correlation other than the very small one due to the finiteness of the 7000-point population; that is, the sampling errors in regression coefficient estimation have approximately equal likelihood of helping or hurting. For positive cross-validation skills more than barely above zero, the sample size–related biases are similar to those using exhaustive, reciprocal sampling as in section 3 (note the close coincidence of the $N = 31$ result in Fig. 6 with the mean $N = 32$ result in Fig. 1b). Recall also the results within the range of positive skill obtained among differing numbers of points held out per development/forecast trial (Fig. 3). Holding out more points at a time in exhaustive, reciprocal cross validation has a form of effect similar to that of eliminating the covariance interdependence in that the magnitude of the negative skills is reduced but the skill behavior a short distance outside the interval in which $|r_{fs}| < r_{crit}$ remains essentially unchanged. Both of these behaviors help provide evidence that the fixed full sample–related degeneracy is concentrated in and near that critical interval.

c. *"Ordinary" hindcast-to-forecast skill decrease*

In the regions of positive cross-validation skill, covariance interdependence under conditions of fixed full sample definition still occurs in amounts as specified by (1). However, as positive skill increases, these effects become negligibly small compared with the other factors causing negative departures from the full sample correlation value (the $y = |x|$ lines in Figs. 1–4). The other factors, depicted in Fig. 6 without a degeneracy, are related to the artificial skill expected in hindcast

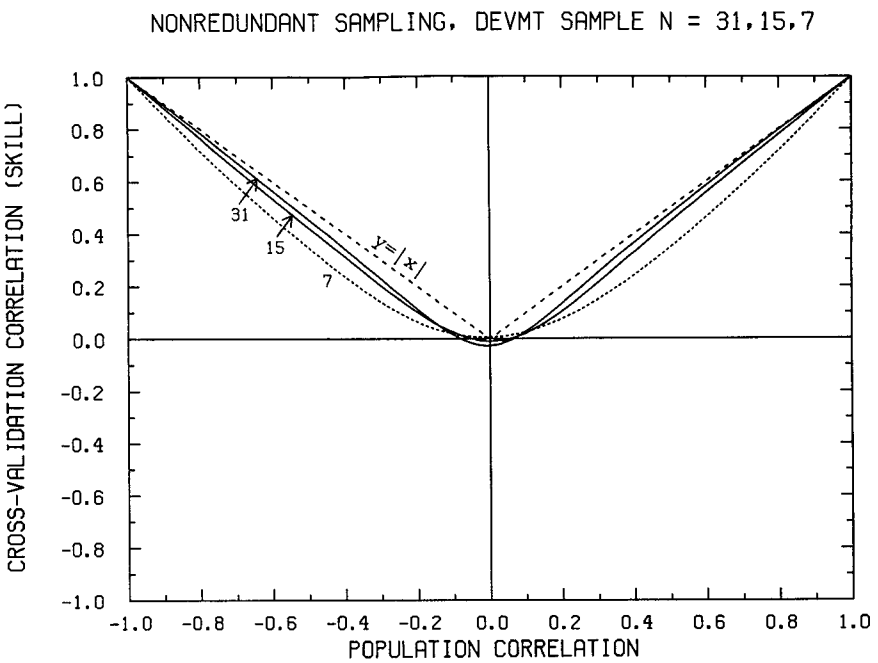

FIG. 6. Cross-validation correlation skill as a function of overall correlation in a 7000-point "population," using 200 nonredundant groups each consisting of N development points and four forecast target points. A dashed $y = |x|$ line and skill curves for $N = 31$, 15, and 7 are shown. See text for further detail.

environments as discussed in Davis (1976) and Michaelsen (1987). Artificial skill arises both because of 1) a positive bias caused by errors of the x and y sample means and variances relative to their population counterparts and 2) correlation sampling variability, in which the relationship between x and y, and thus the regression coefficient, changes by chance among random samples drawn from the same population. The first artificial skill factor is a decreasing function of sample size N, and correlation sampling variability is a decreasing function of both N and r, reflected in the asymmetric confidence interval found on application of the Fisher r-to-z transformation (Brooks and Carruthers 1953). When the predictor and predictand variables are determined prior to the sample selection for the forecast testing procedure (as they are in this study), sampling variability has equal chances for helping or hurting the hindcast skill (i.e., artificial skill may be either positive or negative). It is when the variables are chosen a posteriori on the basis of their predictability within the sample used for the testing (e.g., in dependent sample screening multiple regression or canonical correlation) that sampling variability leads to a positive bias—that is, positive artificial hindcast skill. Whether artificial skill is positive or negative, however, it causes a difference between the sample regression coefficient and the population coefficient, which reduces cross-validation skill. Thus, aside from the degeneracy, negative departures of cross-validation skill from the $y = |x|$ line are caused by random errors in the sample means and variances as well as in the sample x versus y correlation, all of which throw the regression equations off target with respect to the population.

Davis (1976) regards artificial skill as the difference between hindcast skill in a dependent sample and that theoretically possible in the true population. Because the population statistics (and forecast skill) are unknown and the set of statistics obtained from a sample is generally slightly incorrect, the skill realized with cross-validation on independent data (ignoring the degeneracy discussed here) is expected to be still lower than that theoretically possible in the population. Michaelsen (1987) notes that the difference between hindcast skill and cross-validation skill (which he regards as artificial skill—a different definition from that of Davis) is approximately double that between hindcast skill and the unknown theoretically possible population skill (i.e., Davis' artificial skill). In order to maximize cross-validation skills by bringing them closer to the theoretical population skill, Davis (1976) formulated a skill-maximizing damping factor (called K) for sample regression coefficients to be used for forecasts on independent population data.

d. Solutions

There are a number of ways to avoid the interdependent covariance relationship that causes the correlation skill score degeneracy in cross-validation.

Three general types of remedy are possible. They are 1) use of a skill measurement statistic other than the correlation coefficient; 2) a posteriori "cures" for the degenerate scores that occur; and 3) modifications to the cross-validation procedure itself.

The first solution category is the abandonment of the correlation as the skill measurement statistic in favor of alternate parameters such as RSME, mean absolute error, or a categorical skill measure. This choice entails the loss of the desirable features of correlation coefficients, such as their readiness for statistical tests or for conversion to other statistics (like Student's t), or their insensitivity to forecast inflation.

A second category of solution is to apply an arbitrary "quick cure" to the degenerate scores once they occur. This may be justified with the knowledge that degenerate negative scores only appear when the amplitude of the forecasts is very low (resembling climatology forecasts), and a near-zero full sample correlation prevails. The correlation coefficient does not appropriately represent skill expectations when a methodological artifact such as a covariance interdependence dominates the comparatively weak forecast signals reflected in low-amplitude forecasts.

A simple cure is to set all negative correlation scores to zero. This has been done in several studies (Van den Dool 1984; Barnett and Preisendorfer 1987; Barnston and Ropelewski 1992, among others). Another variation of this type of cure is to multiply the negative correlation scores by the ratio of the standard deviation of the set of forecasts to that of the set of corresponding observations over one full round of cross-validation. This would replace the severely degenerate negative biases with small negative biases comparable to those found in the alternate skill score measures.

In a third category of solution the correlation skill measure is used and no a posteriori cures are applied. In this solution the details of the cross-validation procedure are modified to ameliorate the tendency for degenerate scores. There are several options, all of which require using fewer than $N - 1$ points in the development samples.

One option is to keep a constant full sample, but hold out more than one point in each trial. As noted in section 3 (Fig. 3), the degeneracy is weakened as the characteristics of several withheld points are allowed to randomly balance one another. Further improvement would be possible by selecting groups of points to withhold whose deviations from the full sample relationship are designed to be maximally balanced. The more points withheld, the better the potential for designed balancing. However, retaining as large a development sample as possible also remains desirable to reduce sampling variability.

Another way to reduce the degeneracy is to define a series of overlapping "full samples" containing fewer than all N points rather than one N-point sample. A cross-validation exercise across the differing "full samples" can be pooled for verification, making use of the changeable r_{fs} to reduce the inverse covariance interdependence that occurs with rotational one-point removal from a constant full sample.

Still another variant of the same class of remedy is to simulate an operational setting by using only the earliest P percent (e.g., 67%) of the data for the development sample and predicting the next point, then including that one more point in the development sample to predict the next, etc. A drawback in this technique is that the restriction to chronologically forward forecasting greatly reduces the number of forecasts, causing more vulnerability to sampling error in the skill estimate.

The modifications of the cross-validation procedure are time consuming and generally accomplish only a partial remediation of the degeneracy. They also result in greater sampling variability due to the necessity of using smaller development samples. If retention of the correlation skill measure is desired, the a posteriori remedies are clearly the simplest and most effective options, allowing also for use of the maximum $N - 1$ point development samples in the spirit of cross-validation as described in Michaelsen (1987). In view of the mathematical properties of cross-validation as discussed in section 4a, coupled with the implication of the low-amplitude regression forecasts that always accompany degenerate skill scores, setting negative cross-validation correlation skill scores to zero or to weak negative numbers reflecting the forecast amplitude (i.e., the second category of solution) is fully justified and recommended.

Still another approach to a solution is to conduct cross-validation only when the full sample relationship is statistically significant (at least at the 0.05 level, two-sided). When significance is present and predictor(s) is chosen a priori (i.e., not using criteria derived from the sample), the degeneracy typically does not pose a problem and cross-validation results are trustworthy estimates of the lower skill to be expected upon application of the sample regression coefficients to independent data.

All of the above solutions help solve the problem of the occurrence of individual instances of highly negative cross-validation forecast skill. The solutions reduce or eliminate negative "bull's-eyes" on spatial maps of forecast skill and allow for a more representative mean skill score over the spatial domain when that score is computed using the individual correlation skills. It should be noted that such solutions are not needed for a mean score computed without computing the individual point values in the process—that is, by summing over time and space in a large single ensemble computation. In the latter case the degenerate points would have negligible weight in the computation because of their weak forecast amplitudes compared with those of the other points.

5. Summary and conclusions

It has been shown that in predictive skill estimation methods using simple or multiple regression in a cross-validation framework, the skill score based on the correlation between the cross-validation forecasts and the corresponding actual data may be highly negative rather than near zero when actual forecast skill is near zero and the amplitude of the regression forecasts is low.

The degeneracy comes about because when one or more points in a dataset are removed, the statistics of the remaining points are changed in the direction opposite that of the statistics of the removed point(s). When predictability is high and most points exhibit varying degrees of the same predictor–predictand relationship, removing a few points hardly causes a breakdown in the overall statistical relationship. As overall predictability decreases, there is an increasing probability that removal of one or more points may substantively destroy (e.g., change the sign of) the predictive relationship in the remaining development subsample relative to that in the removed points. This can lead to negative skill when averaged over all forecast trials in the cross-validation exercise. When overall predictive skill is close to zero, highly negative cross-validation correlation skill may be expected. In fact, a correlation-based skill score of -1 is producible using experimentally contrived datasets. It is this degeneracy in estimated predictive skill, based on the equal but proportionally opposite anomalous covariance relationship between development points and target point(s), that is described here as a function of the full sample predictor(s) versus predictand correlation for various sample sizes, numbers of points held out, presence of outlier points, and numbers of predictors per forecast.

The negative skill score degeneracy is found to occur in regression-based cross-validation when 1) the forecasts are verified using the correlation between them and the actual data values, and 2) the level of actual (full sample) forecast skill, in terms of correlation magnitude, is at or below approximately $N^{-1/2}$ where N is the full sample size—that is, the two-sided statistical significance level is no stronger than about 0.3. For stronger full sample forecast skill levels the cross-validation correlation skill score is positive and the more "traditional" artificial skill behavior is found. This includes the effects of sampling variability and the degrees of freedom–related biases (Davis 1976; Michaelsen 1987).

There is a choice of several cures to the degeneracy problem. The first is to replace the correlation coefficient as the skill measurement tool with an alternate measure that accounts for the low amplitude of forecasts made in low skill environments, such as RSME or a tally of number of hits in a categorical forecast framework. (In the categorical case this is true only when no forecast inflation is carried out; otherwise, a degeneracy appears.) Either of these would score a set of highly damped, low-amplitude forecasts similar to uniform climatology forecasts. The liability here is the loss of some statistical conveniences associated with the correlation score.

Another option is to use correlation-measured skill but refrain from using nearly all available sample points in the development subsample in each cross-validation forecast trial. This leaves opportunities for a balancing of the peculiarities of points not included in each development sample. "Honest" schemes for selection of maximally balanced sets of points could be designed. The result would be a decrease in the magnitudes of equal-but-opposite deviations from the full sample relationship between development subsample points and withheld forecast target subsample points. Other variations of this general type of remedy exist, such as defining differing overlapping "full samples" made of fewer than N points and cross-validating across many of them in a single set of forecast trials. Unfortunately, all of these methodological modifications are time consuming and generally only weaken rather than eliminate the degeneracy.

Another effective and convenient option is to retain large development subsamples and use correlation verification but to regard all negative skill results as zero. In studies containing spatial distributions of cross-validation correlation skill, the areas of the spatial domains having positive skill estimates are not strongly affected by the degeneracy, so no upward adjustments are needed. A similar option that yields score behavior similar to that of other methods (e.g., RMSE) is to multiply negative correlation skill scores by the ratio of the standard deviation of the forecasts to that of the corresponding observations.

When the predictor–predictand correlation in the full dependent sample is statistically significant at the 0.05 level (two-sided), cross validation is unlikely to produce degenerate skill results. Therefore, another way to safeguard against the degeneracy is to check for full sample significance before applying cross-validation. If significance is achieved, cross-validation can be trusted to provide representative estimates of the lower skill levels to be expected when using the sample regression coefficients to forecast outside the sample. If the full sample correlation is nonsignificant, cross-validation may yield degenerate results and one of the above solutions can be considered.

Provided that the correlation skill score degeneracy is acknowledged and handled appropriately, cross-validation remains quite valid and appealing as a technique to estimate expected predictive skill on an independent dataset.

REFERENCES

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

——, and ——, 1987: Origins and levels of monthly and seasonal

forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.*, **115,** 1825–1850.

Barnston, A. G., and C. F. Ropelewski, 1992: Prediction of ENSO episodes using canonical correlation analysis. *J. Climate,* **5,** 1316–1345.

Brooks, C. E. P., and N. Carruthers, 1953: *Handbook of Statistical Methods in Meteorology,* Her Majesty's Stationery Office, 412 pp.

Davis, R. E., 1976: Predictability of sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.*, **8,** 233–246.

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. Ninth Conf. on Probability and Statistics in Atmospheric Sciences,* Virginia Beach, VA, 175–179.

Huang, J., and H. M. van den Dool, 1993: Monthly precipitation–temperature relations and temperature prediction over the United States. *J. Climate,* 6(6), in press.

Klein, W. H., 1983: Objective specification of monthly mean surface temperature from mean 700-mb height in winter. *Mon. Wea. Rev.*, **111,** 674–691.

Livezey, R. E., and A. G. Barnston, 1988: An operational multifield analog prediction system of United States seasonal temperatures. Part I: System design and winter experiments. *J. Geophys. Res.*, **93,** 10 953–10 974.

——, ——, and B. K. Neumeister, 1990: Mixed analog/persistence prediction of United States seasonal mean temperatures. *Int. J. Clim.*, **10,** 329–340.

Michaelsen, J., 1987: Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.*, **26,** 1589–1600.

Van den Dool, H. M., 1984: Long-lived air temperature anomalies in the midlatitudes forced by the surface. *Mon. Wea. Rev.*, **112,** 555–562.

——, 1987: A bias in skill in forecasts based on analogues and antilogues. *J. Climate Appl. Meteor.*, **26,** 1278–1281.