Cross-Validation in Statistical Climate Forecast Models

JOEL MICHAELSEN

Department of Geography, University of California, Santa Barbara, CA 93106

(Manuscript received 13 August 1986, in final form 25 June 1987)

ABSTRACT

Cross-validation is a statistical procedure that produces an estimate of forecast skill which is less biased than the usual hindcast skill estimates. The cross-validation method systematically deletes one or more cases in a dataset, derives a forecast model from the remaining cases, and tests it on the deleted case or cases. The procedure is nonparametric and can be applied to any automated model building technique. It can also provide important diagnostic information about influential cases in the dataset and the stability of the model. Two experiments were conducted using cross-validation to estimate forecast skill in different predictive models of North Pacific sea surface temperatures (SSTs). The results indicate that bias, or artificial predictability (defined here as the difference between the usual hindcast skill and the forecast skill estimated by cross-validation), increases with each decision—either screening of potential predictors or fixing the value of a coefficient—drawn from the data. Bias introduced by variable screening depends on the size of the pool of potential predictors, while bias produced by fitting coefficients depends on the number of coefficients. The results also indicate that winter SSTs are predictable with a skill of about 20%–25%. Several models were compared. More flexible ones which allow the data to guide the selection of variables generally show poorer skill than the relatively inflexible models where a priori variable selection is used. The cross-validation estimates of artificial skill were compared with estimates derived from other methods. Davis and Chelton's method showed close agreement with the cross-validation results for a priori models. However, Monte Carlo estimates and cross-validation estimates do not agree well in the case of predictor screening models. The results of this study indicate that the amount of artificial skill depends on the amount of true skill, so Monte Carlo techniques which assume no true skill cannot be expected to perform well when there is some true skill.

1. Introduction

One of the most crucial and difficult aspects of developing a statistical forecasting model is obtaining a realistic estimate of forecast skill. As is well known, the standard regression skill estimates based on residual sums of squares are optimistically biased. This bias is commonly referred to in the literature as artificial predictability. The purpose of this paper is to present the results of two small experiments on sea surface temperatures (SSTs) that demonstrate the usefulness of a method for estimating artificial predictability in regression models based on resampling the available data. This technique, known as cross-validation, is a straightforward approach to estimating potential forecast error which is simple to implement, general in its applicability, and nonparametric. In addition, it can provide useful information about specific cases in the dataset which are influential in determining the final forecast model. Cross-validation has intuitive appeal in that it closely mimics the actual forecast situation using data which are similar to those which would be encountered in forecasting. The experiments using cross-validation also provide insight into the relative merits of different model building techniques.

There are two possible sources of bias in the estimate of the predictability inherent in the model development procedure. In all cases, bias is introduced by optimizing the coefficients for the dataset on hand during the model fitting stage. It is relatively easy to estimate this type of bias because it is primarily dependent on the sample size, number of predictor variables, and true skill. Davis (1976, 1977, 1978, 1979) and Chelton (1983) have examined this situation in the case where the predictor set is determined a priori. Their method assumes a large sample size and jointly normal variables. Under these assumptions, artificial predictability is proportional to the number of predictor variables divided by the sample size, and it decreases as true skill increases.

The second possible source of bias arises during the model specification stage if predictor variables are screened based on criteria derived from the sample. This type of bias cannot be estimated by parametric models of the type Davis and Chelton used. Neumann et al. (1977) and Shapiro (1984) have employed Monte Carlo testing to estimate artificial predictability in tropical cyclone prediction models. Rencher and Pun (1980) and Lanzante (1984) carried out sets of Monte Carlo tests to examine false predictability and the inflation of r^2 as a function of sample size, size of the pool of potential predictors, and number of predictors selected. In these studies the central F distribution was used, so the assumptions of normality and no true pre-

dictability are required. Shapiro (1984) suggests that artificial skill decreases with increasing true skill in predictor screening models as it does for a priori selection models. If this is the case, artificial skill estimates derived from Monte Carlo tests will probably be biased upward as a result. Shapiro also points out that forecast skill cannot be related to hindcast skill and artificial skill in a simple manner for screening models as it can in the Davis and Chelton formulations for a priori selection models. Thus, estimates of forecast skill and error are only qualitatively correct.

Cross-validation is used in many forecasting applications in other fields (cf. Breiman et al., 1984). Calling it "jackknifing," Harnack and Lanzante (1985) have applied cross-validation to specification of precipitation. Strictly speaking, jackknifing actually refers to the bias reduction technique proposed by Quenouille (1953) and Tukey (1958) which involves calculating pseudovalues and their means and standard deviations (Stone, 1974). According to Mosteller and Tukey (1977), jackknifing is a general resampling method used to obtain direct estimates of the uncertainty of a statistic. In their example 4 from Chap. 8, jackknifing is used to estimate the uncertainty of the coefficients in a discriminant equation while cross-validation is used to estimate the predictive skill of the equation.

Assessment of the skill of a climatic reconstruction based on calibration with modern instrument data suffers from precisely the same sources of bias as the forecasting case. Thus, cross-validation provides an important tool in this context, too. Lough and Fritts (1985) have employed subsample replication, which is essentially a grouped cross-validation to measure the skill of dendroclimatological reconstructions.

Cross-validation is a generalization of the common technique of omitting a few observations from the model building procedure and then testing the model on the omitted observations. In cross-validation, each observation is left out once and predicted by a model generated from the rest of the data. For a small number of observations, it is probably best to omit each observation one at a time, while larger sample sizes could be handled by omitting a subset each time. Either way, the model is developed on a dataset that has almost as many degrees of freedom as the original, and independent predictions are made for each observation.

It should be noted that in situations where variable screening is used, different equations may result from omitting different observations. In such cases, the whole procedure must be cross-validated, not any specific equation. This is a crucial point because the variable screening phase usually introduces more bias than the coefficient estimation phase.

Cross-validation can also be used as a method for screening a set of potential predictors to find a best subset. The best subset is defined in this instance as the one which minimizes the cross-validation residual sum of squares. This criterion has been called PRESS, or prediction residual error sum of squares (Allen, 1971). If the PRESS criterion is used to select predictor variables, however, it no longer provides an independent estimate of forecast skill. The forecast skill would then have to be estimated through a double cross-validation procedure (Stone, 1974). Each case would be deleted in turn; predictors would be selected by applying the PRESS criterion to the remaining $n - 1$ cases (i.e., deleting each of them in turn) and a prediction would be made for the original deleted case.

The statistical model for cross-validation and some results on its behavior will be described in section 2. Then, as an example, the results of two small experiments using cross-validation will be described. The first attempts to assess artificial predictability of several regression model building techniques applied to the prediction of winter North Pacific sea surface temperatures (SSTs) using only fall SSTs. The second addresses the question of whether or not predictions of winter SSTs can be improved by including fall 700 mb heights and SSTs from summer, spring, and the previous winter. A third section compares the skill estimates obtained using cross-validation with those obtained with other common techniques.

2. Methodology

The history of the development of cross-validation has been well presented by Stone (1974). The idea has been around at least since the 1930s, but refinement of the technique was carried out by a number of authors in the 1960s and 1970s. Two important references from this stage are Stone (1974) and Geisser (1975). Mosteller and Tukey (1977) present a good introduction with examples of both jackknifing and cross-validation. More recently, papers by Stone (1977), Efron (1983), and Bunke and Droge (1984) have used various methods to determine the statistical behavior of cross-validation.

The basic statistical model is as follows. Suppose that one wishes to develop a prediction rule and assess its validity based on a training, or developmental, set $Z = z_1, z_2, \cdots, z_n$ which consists of two parts $z_i = (x_i, y_i)$, where x_i is a vector of observations on potential predictor variables, and y_i is a corresponding observation on a predictand variable. (In this development the predictand is assumed to be a single variable, but it could easily be a vector as well.) Based on the training set, one develops a prediction rule, $\eta(x, Z)$. This rule will then be applied with a future predictor vector, x_0, to arrive at a prediction, $\eta_0 = \eta(x_0, Z)$ of the unobserved predictand y_0.

Let $Q(y_i, \eta_i)$ be a measure of the accuracy of the ith prediction. In regression models estimated by least squares this will usually be $(y_i - \eta_i)^2$. Define the true error rate, Err, as the expected squared error for a randomly selected future case, $z_0 = (x_0, y_0)$; in other words, the expectation

$$\text{Err} = E[Q(y_0, \eta(x_0, Z)].$$

The goal is then to estimate Err on the basis of the training set, Z. The most direct estimate is the resubstitution mean square,

$$\text{mse} = \sum_{i=1}^{n} Q[y_i, \eta(x_i, Z)]/n.$$

In the terminology of climate forecasting, this is the hindcast error. Because this estimate is based on the same data used to construct the rule and the construction process involves some degree of optimization for that set, mse will invariably underestimate Err. In cases with fixed numbers of independent variables chosen before the data are analyzed, this bias can be reduced by reducing the denominator n to reflect the partition of degrees of freedom between the model and the error. If more flexible model building techniques are used, this simple approximation is not available.

Cross-validation approaches the problem of bias by removing each z_i from the dataset in turn, estimating the prediction rule, and then predicting the deleted case. Let $Z_{(i)}$ be the dataset with z_i removed and $\eta(x_i, Z_{(i)})$ be the prediction rule derived from the reduced dataset. Then the cross-validation error estimate is

$$\text{mse}_{(cv)} = \sum_{i=1}^{n} Q[y_i, \eta(x_i, Z_{(i)})]/n.$$

This statistic can be considered to be an estimate of the forecast error. Strictly speaking, the estimate is applicable for sample sizes of $n - 1$, but this is a minor problem for any reasonable sample size greater than about 20.

The cross-validation residuals can be calculated directly from the ordinary least-squares residuals without going through the actual repeated deletion procedure in cases where the pool of predictors in fixed a priori (Belsley et al., 1980). The procedure is as follows: Let $\mathbf{Y}$ be the predictand vector and $\mathbf{X}$ be the matrix with predictor variables in the columns. (The first column will be all 1s if a constant term is included.) Then the hat matrix,

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T,$$

is the matrix which determines the least-squares estimates, $\hat{\mathbf{Y}}$, from $\mathbf{Y}$, i.e.,

$$\mathbf{HY} = \hat{\mathbf{Y}}.$$

Let h_i be the ith diagonal element of $\mathbf{H}$. This is termed the leverage of the ith case and provides a measure of the importance of that case in determining the coefficients of the model.

Consider, for example, the situation where $\mathbf{X}$ is an n by k matrix of normalized EOFs and no constant is included. Then

$$(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{I}$$

and

$$h_i = \sum_{j=1}^{k} x_{ij}^2.$$

So, h_i is the squared distance of the ith case in the predictor space from the mean of all the predictor cases—the origin in this instance.

Now, if $r_i = y_i - \hat{y}_i$ is the ith least-squares residual (calculated with nothing omitted), then

$$r_{(i)} = r_i/(1 - h_i)$$

is the ith cross-validated residual. (Note that this expression is valid in general, not just for orthonormal predictors.) Since $(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ is required to calculate the regression coefficients, very little additional computation is required to get the cross-validation residuals.

This procedure cannot be used when predictor screening is employed. Instead, each case must be deleted in turn, predictor selection and coefficient estimation carried out on the remaining dataset, and a prediction made for the deleted case. Somewhat different sets of predictors could be selected when different cases are deleted, so the cross-validation results will apply to the general model building procedure, rather than to any specific equation.

In general, whether variable screening is used or not, all steps taken in the model development procedure must be repeated in the cross-validation runs. This even includes removing long-term means to obtain anomalies because estimates of the means will change somewhat as each case is omitted.

Cross-validation is straightforward and easy to implement in all cases where a model building rule can be completely computerized. The computational burden may be fairly large for complicated model building procedures applied to large datasets, but computer time is rapidly becoming inexpensive, and the burden is considerably less than that required to carry out extensive Monte Carlo tests. No assumptions were made about the distribution of the data, except that the sample dataset is representative of the population distribution. In addition, it must be assumed that the observations omitted from the model development stage are independent of the set used to derive the model.

In the time-series context, the first assumption means that the time-series must be stationary. There may be some justification for questioning this assumption for a climatic time-series considering that climate does change over long periods. If this is a serious problem, then cross-validation estimates of forecast skill are likely to be optimistically biased. This is the result of a more fundamental shortcoming of regression-based forecast models, however. Regression coefficients are assumed to be constant over time, so the assumption of stationarity is required to justify using regression analysis in the first place. In other words, if the time series are sufficiently stationary to justify using regression analysis to develop a forecast model, then they are sufficiently

stationary to justify using cross-validation to estimate the forecast skill of the model.

The assumption of independence will not be satisfied if there is autocorrelation in the data, and the resulting estimates of forecast skill will be optimistically biased. Most meteorological time series do show some level of autocorrelation, but it is usually small on the time scales of climatic prediction. In the common approach of developing seasonal or monthly forecast equations which are seasonally dependent, each case is separated from the adjacent cases by 1 yr. Few climatic time-series show any significant autocorrelation at lags as long as 1 yr. For example, Madden (1977) and Namias (1978) showed that autocorrelations in surface air temperature 1 yr apart were small except during summer, with the largest values being about 0.4. Even persistent sea surface temperatures have autocorrelation functions which are near zero after 1 yr (e.g., Namias and Cayan, 1981). In situations where moderate levels of autocorrelation present problems, it is possible to insure independence by omitting more than one case at a time and using only the central omitted case. A rough guide to the number of cases which should be omitted each time can be determined using the formula for effective sample size given by Davis (1977) or Chelton (1983). For example, predictor and predictand sets which are both first order autoregressive processes with lag-1 autocorrelations of 0.6 will have an effective sample size of one-half the original sample size. This implies that every other case is independent, so omitting three cases at a time will insure that the central omitted case is independent of the cases remaining in the development set. It should be noted, however, that the reduction in bias obtained by omitting more than one observation is gained at the cost of increased variance. Time series with autocorrelations of 0.25 or less will have an effective sample size which is at least 90% of the original sample size, so the autocorrelation will probably not have any measurable effect on cross-validation estimates of forecast skill.

Studies by Efron (1983) and Bunke and Droge (1984) show that the resubstitution estimate of forecast error has bias of $O(1/n)$. The cross-validation estimate reduces the bias to $O(1/n^2)$. They also show that the standard error of the cross-validation estimate is rather large, although it is smaller than the standard error for the resubstitution estimate. In addition, Efron's results were based on a discriminant model which predicts nominal classes (categories) and has a binary error function (i.e., assigns 1 for a successful prediction, 0 for a failure). He notes that the cross-validation estimate performs better when a smooth residual sum of squares error function is employed.

In general, these results indicate that the cross-validation estimator of false predictability performs markedly better than the resubstitution estimator, especially for standard multiple regression problems. Other more complex estimators may do a better job, but cost is increased considerably and information about the influence of each observation is no longer available. When sample size is large, grouped cross-validation (where several observations are deleted and predicted each time) has better statistical properties, in addition to a reduced computational burden (Bunke and Droge, 1984).

In the following examples, the error produced by always predicting the long-term mean will be referred to as the mean squared anomaly (msa). The hindcast skill will be estimated by the usual percent of variance explained,

$$R^2 = 1 - \text{mse/msa},$$

while the forecast skill will be estimated by the cross-validated percent of variance explained,

$$R_{(cv)}{}^2 = 1 - \text{mse}_{(cv)}/\text{msa}.$$

Artificial skill is the difference between these two measures. Note that this is twice the artificial skill as defined by Davis (1977) and Chelton (1983).

3. Experiments

Data used for the test study were sea surface temperatures (SSTs) from 20 grid points spread across the North Pacific for the period 1947–84. The data were aggregated into seasons. This report will deal only with results for predicting winter SSTs for 1948–84. The first experiment uses cross-validation to test the relative merits of five different model building rules for predicting winter SSTs from fall SSTs. The second utilizes cross-validation to examine the possibility of improving predictions by including SST data from the preceding summer, spring, and winter, along with fall 700 mb heights.

a. Model building rules

Five different model building rules were tested as described herein. The first two involved an a priori model specification. The last three methods used a posteriori variable selection rules to reduce the number of independent variables.

1) *First order autoregressive (AR1):* winter SST anomalies were predicted from fall anomalies at that same grid point.

2) *Prescreened principal components (PC3):* the first three principal components of fall SSTs were selected on the basis of a Monte Carlo test of the significance of the eigenvalues. Note that the selection is a priori and remains fixed throughout the analysis.

3) *Stepwise screening of principal components (PCS):* a pool of the first nine principal components of fall SST anomalies was reduced to a best subset based on the Mallows Cp criterion (Mallows, 1973), which is defined as

$$\text{Cp} = RSS_p/\hat{\sigma} - n + 2(p+1).$$

TABLE 1. Hindcast and forecast skills for the five rules averaged over the whole period 1948–84.

Rule	R^2	$R^2_{(cv)}$	Difference
AR1	0.27	0.21	0.06
PC3	0.34	0.21	0.13
AAS	0.44	0.22	0.22
PCS	0.50	0.14	0.36
PWS	0.62	−0.09	0.71

Here, n is the sample size, p the number of predictors in the subset, RSS_p the residual sum of squares, and $\hat{\sigma}$ an estimate of the error variance which is usually taken to be

$$\hat{\sigma} = RSS_k/(n-k-1),$$

where k is the total number of potential predictors. This rule differs from PC3 in that the selection of components is based on correlations with the dependent variables, rather than on variance of the independent dataset.

4) *Stepwise screening of grid points (PWS):* all of the 20 fall grid point series were run through a best subsets selection procedure based on Mallows Cp. This technique is commonly called point-wise screening.

5) *Area averaged screening (AAS):* correlations between each of the 20 fall grid point series and the winter grid point series were used to identify coherent regions in the predictor field. A coherent region was defined as one or more adjacent grid points which had correlations with the predictand that were both significant and of the same sign. All the grid point series within a region were averaged together to form a single predictor for each region. This technique was suggested by Barnett (personal communication, 1984).

Table 1 gives a summary of the results for the five model-building rules averaged over all 20 grid points. The hindcast skills for the two a priori selection rules show the least inflation over the forecast skills. The AR1 rule with its single parameter shows the lowest hindcast skill but also the lowest false skill, making its performance about equal to that of the PC3 rule. The latter has a higher hindcast skill and also higher false skill because it had three parameters to estimate.

Clearly, variable screening enhances the opportunity for including false predictability in the resulting equations. The two stepwise screening models show the highest hindcast skills, but the lowest forecast skills. The PWS rule performs particularly poorly, with its impressive hindcast skill disappearing completely in the cross-validation. The somewhat better performance of the PCS rule is apparently a result of the smaller pool of potential predictors. The AAS rule, on the other hand, shows forecast skill comparable to the two a priori screening rules, with the higher false skill being offset by higher hindcast skill. Apparently, the areal averaging produces more stable sets of predictors than the use of individual grid points in the PWS rule. It is possible that the poor performance of the screening models results from the choice of the stopping rule. As Mallows points out, however, the Cp criterion is closely related to the other commonly used criteria based on adjusted r^2, F statistics, or t statistics.

The poor overall performance of the PWS rule is mainly the result of large negative skills for several grid points which had low forecast skills for all rules. For example, 25°N/155°E had a top forecast skill of 0.10 for the PC3 rule. For the PWS rule, the hindcast skill was 0.70, and the forecast skill was −0.75. On the other hand, the ten grid points which showed the highest skill for the PC3 rule, all over 0.20, showed an average forecast skill for the PWS rule of 0.23.

The general result seems to be that larger levels of forecast skill can be identified reasonably well by any technique, but in areas of low skill the last three, more flexible rules will focus on sample-specific noise resulting in spectacularly poor forecast skills. In the worst cases, both the PCS and PWS methods showed forecast errors greater than twice the original variance—25% greater for the AAS and PC3 methods and 15% greater for the AR1 method.

One of the complaints about using prescreened principal components rules like the PC3 rule has been that the predictors are selected based on a criterion which maximizes the variance contained in the predictor set rather than the covariance between the predictor and predictand sets. With this in mind it is interesting to note the number of times lesser components were selected in the PCS method. The results are as follows: the first component was selected at 11 of the 20 grid points; the second at 14; the third at 5; the fourth at 8; the fifth at 9; the sixth at 3; the seventh at 8; the eighth at 5; and the ninth at 7. The first two components are the most useful by this measure, but there is not a consistent drop in importance beyond that. This result seems to suggest that there is useful information in the lesser components. The PC3 method showed better forecast skill than the PCS method, though, so additional information in the lesser components is apparently difficult to separate from the additional noise.

The AR1, PC3, and AAS methods all show approximately the same level of forecast skill. Looking more closely at their temporal behavior, Fig. 1 shows the msa and forecast errors ($mse_{(cv)}$) of the three rules for each winter. The similarity of all the curves indicates that years with large msa will also have large forecast errors and that all three methods are performing comparably on a year-by-year basis. The msa for the winter of 1949 stands out well above the rest, suggesting that some of the early data might be suspect. The anomaly pattern (not shown) is quite similar to ones in later years, but its magnitude is unusually large. The early 1950s and most of the 1960s stand out as periods with relatively small anomalies.

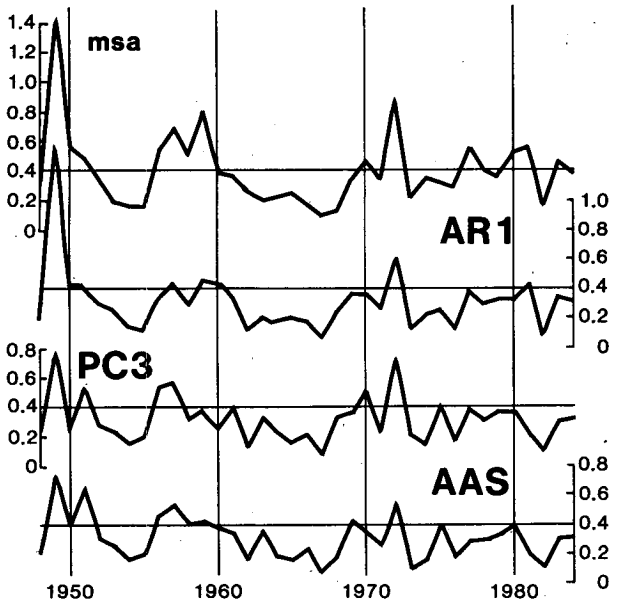

FIG. 1. The msa (top) and absolute forecast errors, $mse_{(cv)}$, for the AR1, PC3, and AAS rules averaged over all 20 grid points.

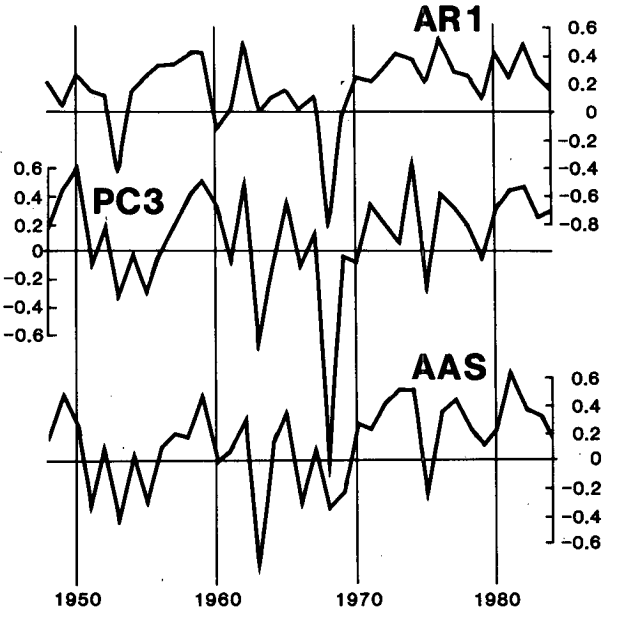

FIG. 2. The relative forecast skills, $r^2_{(cv)}$, for the AR1, PC3, and AAS rules averaged over all 20 grid points.

Comparing Fig. 1 with the relative forecast skills for the three rules plotted in Fig. 2 indicates that the years with large anomalies will have large absolute errors but also large relative skills. The years with small anomalies, on the other hand, will tend to have small absolute errors, but also small or negative relative skills. For example, the average forecast errors for the 10 yr with the largest mean squared anomalies are 0.54 for the AR1 rule, 0.49 for the PC3 rule, and 0.44 for the AAS rule. The average errors for the 10 yr with the smallest anomalies are 0.17, 0.22, and 0.19. The average relative skills for the 10 large anomaly years are 0.26, 0.29, and 0.34, while the comparable figures for the 10 small anomaly years are −0.02, −0.30, and −0.10. The AR1 rule again appears to be the most conservative, with a lesser tendency for negative skills during small anomaly years and also a lesser tendency for large positive skills during large anomaly years.

Figure 3 shows the difference between hindcast and forecast error for the three rules. As expected, the forecast error is larger in all cases. The largest differences typically occur during years with large anomalies indicating that these years are the most influential in determining the predictive equations. The winter of 1949 stands out along with 1951 and 1972 for the AAS and PC3 rules. In spite of the differences between hindcast and forecast errors for large anomaly years, however, skills are still generally best during these years.

It is interesting to note that there appear to be long-term trends in the forecast skills. As can be seen in Table 2, all three rules perform poorly in the 1960s, a period characterized by small amplitude anomalies. The poor performance of the AR1 rule and the good performance of the PC3 model during the first period (1948–53) are caused primarily by their contrasting skills at predicting the large anomalies in 1949. The AR1 rule performs exceptionally well during the late 1950s which were characterized by an unusually orderly progression of large anomalies from west to east across the midlatitude ocean as noted by Favorite and McLain (1973) and many others. The improvements since 1970

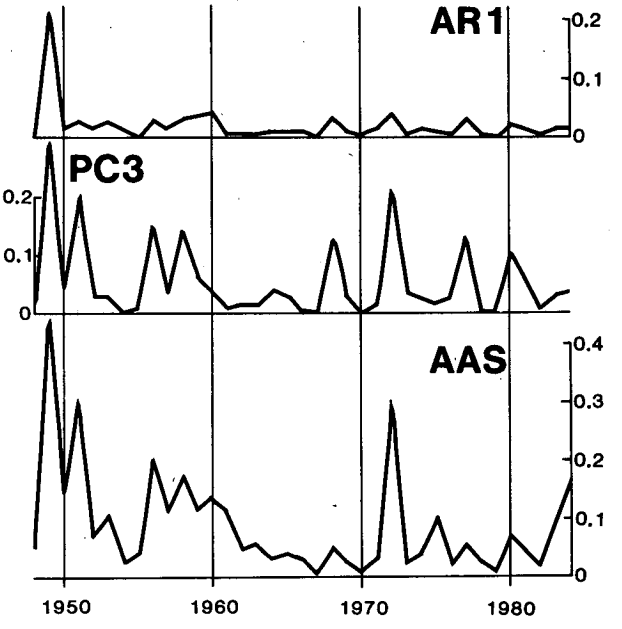

FIG. 3. The artificial skills, $mse_{(cv)}$ − mse, for the AR1, PC3, and AAS rules averaged over all 20 grid points.

TABLE 2. Mean square anomaly and forecast and hindcast skills averaged over six year periods.

		AR1		PC3		AAS	
Years	msa	R^2	$R^2_{(cv)}$	R^2	$R^2_{(cv)}$	R^2	$R^2_{(cv)}$
1948–53	0.55	0.18	0.09	0.49	0.31	0.56	0.22
1954–59	0.47	0.41	0.38	0.37	0.23	0.46	0.22
1960–65	0.28	0.13	0.09	0.19	0.10	0.28	0.03
1966–71	0.26	0.10	0.05	0.03	−0.08	0.12	0.02
1972–77	0.44	0.38	0.34	0.40	0.23	0.56	0.35
1978–84	0.40	0.29	0.26	0.36	0.28	0.46	0.31

in the AR1 rule and particularly in the AAS rule lead one to wonder if data quality has improved because anomaly magnitudes were not uniformly large.

b. *Expanding the predictor set*

This section presents the results of three different methods for including information from earlier seasons and from fall 700 mb heights in the winter prediction model. The three methods were as follows.

1) *Second, third, and fourth order autoregressive (AR2, AR3, AR4):* Winter SSTs were predicted by SSTs from summer and fall at the same location (AR2); from spring, summer, and fall at AR3; and from the previous winter, spring, summer, and fall at AR4. In all cases the predictor sets were fixed.

2) *Screened autoregressive (ARS):* The best predictor set was selected from the pool consisting of SSTs from fall, summer, spring, and the preceding winter at the same location. The selection criterion was minimization of the cross-validated sum of squares, or PRESS. A second level of cross-validation was required to obtain an independent estimate of forecast skill.

3) *Principal components field screening (PCFS):* The principal components of fall SSTs and 700 mb heights, summer SSTs, spring SSTs, and SSTs from the preceding winter were computed. For each field the number of components to retain was determined by Monte Carlo testing. Then the best set of fields was selected by the same PRESS criterion used in method 2. Once again, double cross-validation was used to obtain estimates of forecast error. Note that screening was carried out on fields, not on individual components of the fields. Thus, all the preselected components were either retained or discarded resulting in a reduction in the number of decisions made by the data and, possibly, the artificial skill.

The results for the three methods are summarized in Table 3. The first point to note is that none of the models perform better than the models which included only fall SSTs. There is a suggestion that the best autoregressive model might include fall and spring SSTs but not summer SSTs. The tendency for autocorrelations to drop in summer was first noticed by Namias and Born (1970) and has often been attributed to the formation of shallow anomalies during the season of weak vertical mixing.

Another point to note is that the performance of all the methods is roughly comparable. This suggests that there is a limited amount of information in the predictor fields which is simply being accessed in different forms by the different models. An alternate interpretation is that there is a limit to the amount of information which can be discriminated from a dataset with only 37 degrees of freedom. If, for example, 80 years of data were available, it might be possible to include more useful information from preceding seasons without exhausting the degrees of freedom and enlarging the artificial skill.

In the autoregressive screening model, fall was selected at 17 of the 20 grid points, summer at 5, spring at 7, and the preceding winter at 5. This reinforces the idea that fall is by far the most important season for predicting winter and that spring is slightly more useful than summer. Another indication of the importance of fall SSTs is provided by the frequencies of selection for the PCFS method. Fall SST PCs were chosen 15 times; fall 700 mb PCs, 7 times; summer SST PCs, 3 times; spring SST PCs, 3 times; and the preceding winter SST PCs, 4 times.

The relatively high artificial skills of the screening methods in both experiments demonstrates the importance of the model selection phase in introducing error. For fixed predictor models the sole source of artificial skill is the optimization of the coefficients, while the screening models add the artificial skill involved in selecting predictors. It is possible to get an idea of the relative size of these two sources by estimating what the forecast skill would have been if the predictor set selected by a screening procedure had been chosen and fixed a priori. Table 4 shows the total artificial skill and the contributions from the coefficient fitting and predictor selecting phases, along with the average number of coefficients estimated and the size of the predictor pool for the five screening models. The fitting error is directly related to the number of coefficients estimated while the selection error is related to the size of the potential predictor pool—in other words, to the number of decisions required in the selection process.

With these principles in mind there are several interesting points to be made about the results in Table

TABLE 3. Hindcast skills, forecast skills, and artificial skills for expanded predictor set methods.

Rule	R^2	$R^2_{(cv)}$	Difference
AR2	0.29	0.19	0.10
AR3	0.34	0.20	0.14
AR4	0.36	0.18	0.18
ARS	0.32	0.18	0.14
PCFS	0.44	0.18	0.26

TABLE 4. Total artificial skill and components produced by estimating coefficients and selecting predictors for the screening rules. Also included are the average number of predictors selected for the 20 grid points and the size of the predictor pool.

Rule	Total artificial skill	Estimation component	Selection component	Average number of coefficients	Size of predictor pool
AAS	0.22	0.06	0.16	1.8	20
PCS	0.36	0.14	0.22	3.5	9
PWS	0.71	0.14	0.57	5.0	20
ARS	0.14	0.06	0.08	1.8	4
PCFS	0.26	0.14	0.12	4.8	5

4. In general the selection error is larger than the fitting error. This is particularly true for the AAS and PWS methods which start with a pool of 20 potential predictors but end with a much smaller number of coefficients to be estimated. One of the advantages of the AAS method is that area averaging means only one parameter is estimated for each area, rather than one for each grid point. The PCFS results run counter to the general trend in that the fitting component is larger than the selection component. This is apparently due to the grouped screening where selection decisions were to either retain or omit the whole field, but where more than one coefficient must be estimated for each retained field. The most conservative method, ARS, again showed the lowest artificial skill particularly in the screening phase.

c. *Comparisons with other techniques*

The cross-validation estimates of artificial predictability for the SST models discussed before were then compared with estimates calculated using the two main alternative techniques employed in meteorological prediction work. Using the Davis/Chelton technique, artificial predictability was estimated for the a priori selection models. Results for the screening selection models were compared to results determined by Monte Carlo simulations.

The Davis/Chelton method requires a large sample size and normally distributed data. A sample size of 37 is reasonable for methods with accuracies of $O(1/n)$, and the SST data are not markedly nonnormal, so this approach is appropriate. The first task is to determine the effective sample size. The procedure of determining long-lag hindcast skills was used, and results showed that there are close to $n - 1$ effective degrees of freedom. This is to be expected, because the correlation between SST anomalies 1 yr apart is usually small. Chelton's (1983) Eq. (15) was used with $n^* = n - 1$ to estimate the artificial skills, keeping in mind that artificial skill as defined in this paper is twice that defined by Chelton. The results for all 20 grid points are as follows: AR1—0.04; PC3—0.12; AR2—0.08; AR3—0.12; AR4—0.16. Comparing these figures with the difference columns in Tables 1 and 3, it is clear that the Davis/Chelton method and cross-validation produce very similar results. In all cases the Davis/Chelton estimates are slightly smaller, but all the differences are 0.01 or 0.02, well within $O(1/n)$.

The results for individual grid points are also quite comparable with one exception. The point at 45°N/175°E shows cross-validation artificial skills which are larger than the Davis/Chelton estimates for all methods by between 0.10 and 0.16. The winter of 1949, which has already been noted as unusual, is a particularly large outlier at this grid point, and it has a great impact on the values of the coefficients. As a result, the cross-validation residual is much larger than the already large least-squares residual. Almost two-thirds of the artificial predictability is produced by this 1 yr. The presence of this outlier apparently presents serious problems for the assumption of normality in this one case, and also highlights the utility of cross-validation as a tool for identifying outliers.

Estimates produced by the Monte Carlo technique employed by Shapiro (1984), Lanzante (1984), and others were compared to the cross-validated forecast skills for the ARS, PCS, and PWS methods. The predictand variable was randomly shuffled 100 times to produce models with no true skill. The number of predictors selected for the original models was held fixed, and twice the average hindcast skill was used to estimate artificial skill as defined in this paper. (The factor of 2 arises because artificial skill is defined in this paper as the difference between hindcast and forecast skill while Shapiro 1984, defines it as the difference between hindcast skill and true skill and assumes that the difference between true skill and forecast skill is comparable.)

The average artificial skills estimated by Monte Carlo methods are: ARS, 0.17; PCS, 0.41; and PWS, 0.71. The cross-validation estimates from Tables 1 and 3 are ARS, 0.14; PCS, 0.36; and PWS, 0.71; so on average, the two methods produce comparable results. There is a considerable amount of divergence for individual grid points, however, particularly for the PCS and PWS models. The cross-validation estimates of artificial skill exceed the Monte Carlo estimates by at least 0.25 at four grid points for the PCS model and at three grid points for the PWS model, while the Monte Carlo estimates exceed the cross-validation estimates by at least 0.25 at four grid points for the PCS model and at five grid points for the PWS model.

In general, these discrepancies appear to be related

to the stability of the screening process under cross-validation. Grid points where the cross-validation artificial skill estimates were smaller than the Monte Carlo estimates usually showed relatively small variations in the predictors selected in the cross-validation runs compared to those selected from the original dataset. Conversely, grid points where cross-validation estimates were larger than Monte Carlo estimates showed relatively large variations in the set of predictors selected in cross-validation runs.

To quantify this idea, consider a measure of the stability of the screening process which counts the number of differences between predictors selected in a cross-validation run and those selected for the main equation. In other words, define a variable, $k_j, j = 1, \cdots J$, which equals 1 if the kth predictor is selected for the main equation and 0 otherwise; J is the size of the predictor pool. Define a similar variable $k_{(i)j}$ for the equation derived with the ith case deleted. Then the mean number of differences is

$$\bar{d} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{J} |k_j - k_{(i)j}|.$$

This measure ranges between 0 (if the same set of predictors is selected in all the cross-validation runs as in the main equation) and J (if the cross-validation equations select none of the predictors from the main equation and select all of the predictors omitted from the main equation).

For the PCS method, $\bar{d}$ averages 0.17 for grid points where the cross-validation estimates were at least 0.05 smaller than the Monte Carlo estimates and 0.82 for grid points where the cross-validation estimates were at least 0.05 larger. The comparable averages for the PWS method are 0.88 and 2.74.

These figures illustrate an important difference between the Monte Carlo and cross-validation approaches. The former fixes the number of predictors to the number selected for the original equation. Davis (1977) and Shapiro (1984) have pointed out that this is necessary; otherwise a smaller number of predictors is likely to be selected in the Monte Carlo runs when true skill is zero resulting in an underestimate of artificial skill. A consequence of this fact is that the selection criterion used to determine the number of predictors in the original equation cannot be used in the Monte Carlo runs. In both the PCS and PWS models the pool of predictors was the same for all grid points, so the artificial skill estimates depend almost entirely on the number of predictors selected for the main equation.

Cross-validation, on the other hand, fixes the selection criterion, not the number of predictors selected. As a result, it is possible to select different numbers of predictors in cross-validation runs. Changes in the number of predictors selected helps to increase $\bar{d}$ and generally leads to higher cross-validation estimates of artificial skill. Thus cross-validation permits an assessment of the stability of the number of predictors selected.

A second cause of large values of $\bar{d}$ is changes in the ordering of the predictor set in cross-validation runs which results in one predictor being substituted for another. Shapiro (1984) notes that the existence of some skill will order the predictors, while the ordering in the Monte Carlo runs will be random. He suggests that this ordering will tend to reduce the artificial skill below the estimates for random ordering. Results of this analysis do show that the ordering of predictors in the Monte Carlo runs is random. Each potential predictor is selected with approximately equal frequency. Since cross-validation, unlike the Monte Carlo approach, does preserve most of the correlation structure between the predictor pool and the predictand, the ordering in the original equation does tend to be preserved to a degree in the cross-validation runs. Grid points with low values of $\bar{d}$ generally show strong ordering, so the same set of predictors is selected most of the time. The lower estimated artificial skills for these points tend to support Shapiro's contention that the Monte Carlo approach overestimates artificial skill for models with a strong ordering of predictors. The high cross-validation artificial skills for grid points where the ordering of variables is weak or unstable suggests, however, that in these cases Monte Carlo artificial skill estimates can be too low. This results from the fact that changes in predictors selected will not have a detrimental effect in situations where there is no true skill and, therefore, no ordering of predictors. On the other hand, if there is true skill but the ordering of predictors is weak or unstable, then changes in the predictors selected will frequently occur in cross-validation runs, and forecast skills will be degraded as a result.

In general, there are two factors which determine the stability of the ordering of predictors. The first is the correlation structure between the predictors and predictand. If a small number of predictors have much higher correlations with the predictand than the rest of the predictors, then they will be chosen consistently in cross-validation runs. This situation becomes more likely as the true skill increases. If, on the other hand, true skill is weak or there are many predictors which have comparable levels of correlation with the predictand, there will be many changes in the set of predictors selected in the cross-validation runs, and estimated artificial skills will be large. The second factor is the presence of outliers. Specifically, outliers occurring in both the predictand and predictor sets for the same case will produce major changes in the selection of predictors when that case is deleted. The impact of extreme values was noted for the a priori selection models, but this problem is potentially much more serious for screening models. Outliers will have a strong impact on both the selection of predictors and the estimation of coefficients, producing much more opportunity for artificial skill.

Both of the above factors depend on the correspondence between the predictand and predictor set. This correspondence will not be captured in a Monte Carlo test, even one which incorporates nonzero true skill, so neither factor can be effectively measured.

The comparison of cross-validation to Monte Carlo testing highlights some of the difficulties in estimating artificial skill in screening models. On average the two techniques produce comparable results, but the large discrepancies at individual grid points are bothersome. Part of the problem may result from the high variance of cross-validation estimates, but it appears that the Monte Carlo technique does not adequately measure the impact of the stability of the selection process and the strength of ordering in the predictor set. Cross-validation does measure this effect, and the results of the comparison with Monte Carlo testing suggest that strong ordering will reduce artificial skill below the value for no true skill, as suggested by Shapiro, but also that weak ordering can increase the artificial skill above the value for no true skill.

4. Summary

The results of these experiments using cross-validation reinforce the point that knowledge gained from the model-building dataset is biased. True forecast skill may or may not increase, but artificial skill is bound to increase as more information from the model-building dataset is employed in developing the model. This bias, or artificial predictability, increases with each decision from the data. Bias introduced by variable screening depends on the size of the pool of potential predictors, while bias produced by fitting coefficients depends on the number of variables retained. In most of the examples given herein, bias from variable selection appears larger, although it can be offset somewhat by low bias from coefficient estimation if a small number of variables is chosen. Fixed predictor models do not suffer from selection bias, but one must choose between including a large number of predictors which will produce a high estimation bias and omitting predictors which might contain legitimate information. The latter choice appears to be wiser in situations where degrees of freedom are scarce. This implies that selecting predictors a priori by some more or less arbitrary criterion which does not make use of the covariance between the predictor and predictand sets may be a good idea. In this context, principal components or some similar dimension reduction technique which only operates on the predictor set can be very useful as the a priori criterion.

Comparisons of cross-validation with other methods of estimating artificial skill indicate another advantage of a priori selection methods; it is much easier to obtain an honest estimate of artificial skill. Cross-validation estimates show close agreement with estimates derived using the technique of Davis (1979) and Chelton (1983), so either technique should produce acceptable results. The advantage of cross-validation in this instance, however, is its ability to identify and assess the effect of outliers and other deviations from normality. Furthermore, cross-validation residuals can be obtained with minimal additional computation.

The comparison of cross-validation in predictor screening models with the Monte Carlo methods employed by Neumann et al. (1977), Lanzante (1984), Shapiro (1984), and others emphasizes the fact that screening makes estimation of artificial skill much more difficult. The basic problem is that the behavior of artificial skill in situations where there is nonzero true skill. The results of this study suggest that the stability of the selection process is important in determining the amount of bias introduced by predictor screening. Predictor selection is random in the absence of true skill, so stability is not an issue. As Shapiro points out, however, true skill introduces an ordering in the selection process, and it appears that the stability of that ordering is an important factor. Cross-validation artificial skill estimates are generally smaller at grid points where the screening process is stable (in the sense that predictors selected in most of the cross-validation runs are the same as those in the original equation). In this case, the presence of nonzero true skill and a strong ordering of the predictors reduces the bias produced by the selection process below what it would be for a model with no skill, as noted by Shapiro (1984). On the other hand, cross-validation artificial skill estimates are higher than Monte Carlo estimates at grid points where there are large variations in the predictors selected in the cross-validation runs. This result suggests that the presence of weak or unstable ordering of predictors can create larger selection bias than purely random ordering. In the latter case, substitutions of one predictor for another do not make much difference, but in the former case, substitutions can have an important impact on artificial skill. Unstable ordering is caused by several predictors having approximately the same level of correlation with the predictand. It is magnified by the presence of outliers in both sets which have a major impact on the size of the correlations and, therefore, on the ordering of the predictors.

Monte Carlo techniques based on zero true skill cannot be expected to perform well if there is true skill. It would be better to design a Monte Carlo test which built in some assumed level of true skill, but even this approach would not preserve the correlation structure between the predictor pool and the predictand and, therefore, would not accurately measure the amount of bias introduced by the selection process in the presence of nonzero true skill. Furthermore, Monte Carlo estimates do not necessarily represent upper bounds on artificial skill as suggested by Shapiro. It is possible that the relatively high variance of the cross-validation estimator overemphasizes the importance of stability or instability in the selection process, but at least it treats the problem explicitly. Thus, cross-validation

would seem to be the preferred method, but further studies are needed to verify this assertion.

In general, it appears that the best strategy for developing a statistical forecasting model will depend on the amount of true skill. In situations where relatively high true skill is anticipated—specifying surface temperatures from 700 mb heights, for example—screening methods will probably be useful tools for picking out a small set of important predictors. In most actual climate forecasting applications, however, true skill is relatively small, and an a priori selection method appears more likely than screening methods to produce a reasonable set of predictors. It is difficult to generalize about how much skill is required to support a variable screening method, but a rough guideline can be obtained by noting that the screening methods generally produced results comparable to those obtained by the a priori selection methods at grid points where forecast skills were at least 25%–30%.

While the experiments discussed above were not designed to be full tests of either the relative merits of the model building rules or of the predictability of North Pacific SSTs, some interesting points emerged which deserve emphasis.

1) The poor performance of the PWS rule is in direct contradiction to the findings of Klein and Walsh (1983) for specification of United States temperatures. They did not carry out a full cross-validation, but they did do limited tests on an independent dataset. Assuming their results did hold up after full cross-validation, there are two possible explanations for this contradiction. First, they had many more degrees of freedom with which to support the more flexible model. Second, as noted above, the PWS method did a reasonably good job in cases where the forecast skill, as measured by other rules, was high. In their example of specifying temperature from 700 mb heights, a high degree of skill could be expected on physical grounds.

2) Neither cross-validation nor the other techniques measure the judgment of the researcher in model building. It is possible, for example, that a point-wise screening technique could be quite effective if the researcher used knowledge and experience to guide the variable selection process instead of blindly following output from a computerized routine.

3) Estimates of forecast skill based on small subsets of the data show a large amount of variability. Therefore, the common practice of reserving a few cases from the end of the data and using them to estimate forecast skill can give misleading results. A full cross-validation will clearly give more stable estimates.

4) Double cross-validation does provide a way to use estimated forecast skill as a criterion for selecting variables and still obtain an independent estimate of the forecast skill. The computational burden is increased markedly, however, and it seems likely that a selection criterion based more directly on the mse (e.g., adjusted r^2 or Cp) would perform almost as well with a considerable saving in computer time. In general, the choice of variable selection criteria does not appear to be an issue of major importance.

5) The comparable forecast skills for the AR1, PC3, and AAS rules suggest that similar amounts of information are being captured by each. However, they show marked differences in performance in individual years. It might be useful to employ the most conservative model in years when the forecasted anomalies were small and a more flexible model in years when the forecasted anomalies were large. If this approach were used, it would be necessary to include the decision of which model to employ in the cross-validation runs.

6) The results indicate that one should be hesitant to include early data of questionable quality. Its inclusion could produce important changes in the model form and coefficient values or produce overly inflated estimates of forecast error.

7) In addition to providing a means for estimating forecast skill, cross-validation can also highlight influential cases or outliers—either the result of bad data or of legitimate extreme conditions. In the latter case cross-validation results can focus attention on specific years for more careful case studies.

8) Results just cited can be applied to climatic reconstruction techniques, as well as forecasting.

9) There appears to be an average skill of about 25% for predicting winter SST data used in this study. In some areas along the coast of North America and in the midlatitude ocean near the dateline the figure approaches 50%. Most of the predictive skill resides in the fall SSTs, although a couple of grid points showed marked improvement when fall 700 mb heights were included. It appears that just about any fairly conservative model building technique will capture most of the predictability. It is possible that SSTs or 700 mb heights from other seasons might turn out to be important if one had more degrees of freedom to work with, but the current dataset did not provide enough flexibility to screen the additional predictive skill from the sample-dependent noise.

Acknowledgments. The author wishes to thank Dr. Tim Barnett for valuable discussions, the anonymous reviewers for useful suggestions, and Mr. David Lawson for preparing the final figures. The research was funded in part by the California Space Institute.

REFERENCES

Allen, D. M., 1971: Mean square error of prediction as a criterion for selecting variables. *Technometrics,* **13,** 469–475.

Belsley, D. A., E. Kuh and R. E. Welsch, 1980: *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley, 292 pp.

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone, 1984: *Classification and Regression Trees.* Wadsworth, 358 pp.

Bunke, O., and B. Droge, 1984: Bootstrap and cross-validation estimates of the prediction error for linear regression models. *Ann. Stat.,* **12,** 1400–1424.

Chelton, D. B., 1983: Effects of sampling errors in statistical estimation. *Deep-Sea Res.*, **30,** 1083–1103.

Davis, R., 1976: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.*, **6,** 249–266.

——, 1977: Techniques for statistical analysis and prediction of geophysical fluid systems. *Geophys. Astrophys. Fluid Dyn.*, **8,** 245–277.

——, 1978: Predictability of sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.*, **8,** 233–246.

——, 1979: A search for short range climate predictability. *Dyn. Atmos. Oceans,* **3,** 485–497.

Efron, B., 1983: Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.*, **78,** 316–331.

Favorite, F., and D. R. McLain, 1973: Coherence in trans-Pacific movements of positive and negative anomalies of sea surface temperature, 1953–60. *Nature,* **244,** 139–143.

Geisser, S., 1975: The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, **70,** 320–328.

Harnack, R. P., and J. R. Lanzante, 1985: Specification of United States seasonal precipitation. *Mon. Wea. Rev.*, **113,** 319–325.

Klein, W. H., and J. E. Walsh, 1983: A comparison of pointwise screening and empirical orthogonal functions in specifying monthly surface temperatures from 700 mb data. *Mon. Wea. Rev.*, **111,** 669–673.

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.

Lough, J. M., and H. C. Fritts, 1985: The Southern Oscillation and tree rings: 1600–1961. *J. Climate Appl. Meteor.*, **24,** 952–966.

Madden, R. A., 1977: Estimates of autocorrelations of seasonal mean temperatures over North America. *Mon. Wea. Rev.*, **105,** 9–18.

Mallows, C. L., 1973: Some comments on Cp. *Technometrics,* **15,** 661–675.

Mosteller, F., and J. W. Tukey, 1977: *Data Analysis and Regression.* Addison-Wesley, 586 pp.

Namias, J., 1978: Persistence of U.S. seasonal temperatures up to one year. *Mon. Wea. Rev.*, **106,** 1557–1567.

——, and R. M. Born, 1970: Temporal coherence in North Pacific sea surface temperature patterns. *J. Geophys. Res.*, **75,** 5952–5955.

——, and D. R. Cayan, 1981: Large-scale air–sea interactions and short-period climatic fluctuations. *Science,* **214,** 869–876.

Neumann, C. J., M. B. Lawrence and E. L. Caso, 1977: Monte Carlo significance testing as applied to statistical tropical cyclone prediction models. *J. Appl. Meteor.*, **16,** 1165–1174.

Quenouille, M. H., 1953: Notes on bias in estimation. *Biometrika,* **43,** 353–360.

Rencher, A. C., and F. C. Pun, 1980: Inflation of $R sup 2$ in best subset regression. *Technometrics,* **22,** 49–53.

Shapiro, L. J., 1984: Sampling errors in statistical models of tropical cyclone motion: A comparison of predictor screening and EOF techniques. *Mon. Wea. Rev.*, **112,** 1378–1388.

Stone, M., 1974: Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc.* **B36,** 111–47.

——, 1977: Asymptotics for and against cross-validation. *Biometrika,* **64,** 29–38.

Tukey, J. W., 1958: Bias and confidence in not-quite large samples. *Ann. Math. Statist.*, **29,** 614.