Artificial Skill and Validation in Meteorological Forecasting

PAUL W. MIELKE JR.

Department of Statistics, Colorado State University, Fort Collins, Colorado

KENNETH J. BERRY

Department of Sociology, Colorado State University, Fort Collins, Colorado

CHRISTOPHER W. LANDSEA AND WILLIAM M. GRAY

Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado

(Manuscript received 21 February 1995, in final form 8 November 1995)

ABSTRACT

The results of a simulation study of multiple regression prediction models for meteorological forecasting are reported. The effects of sample size, amount, and severity of nonrepresentative data in the population, inclusion of noninformative predictors, and least (sum of) absolute deviations (LAD) and least (sum of) squared deviations (LSD) regression models are examined on five populations constructed from meteorological data. Artificial skill is shown to be a product of small sample size, LSD regression, and nonrepresentative data. Validation of sample results is examined, and LAD regression is found to be superior to LSD regression when sample size is small and nonrepresentative data are present.

1. Introduction

Recently developed prediction models of various atmospheric phenomena have motivated this study (Gray et al. 1992, 1993, 1994). We are interested in the influence of various conditions on the degree of agreement between observed values and values predicted by a meteorological regression model. Of particular interest are differences between least (sum of) absolute deviations (LAD) regression models and least (sum of) squared deviations (LSD) regression models (commonly termed least squares models) under a variety of research conditions. Such conditions include sample size, the inclusion of uninformative independent variables, and the influence of the amount and severity of nonrepresentative data.

In the context of artificial skill and validation in meteorological forecasting, we present examples of prediction of intensity change of Atlantic tropical cyclones 24 h into the future. Datasets containing values of independent variables that deviate either moderately or severely from the bulk of the available data are termed ''nonrepresentative'' or ''contaminated'' datasets. The use of contaminated data is to represent potential situations that frequently arise in forecasting, whether it be day-to-day or seasonal forecasting. For predictions of tropical cyclone intensity change in this case, large errors (or contamination) in the predictors can frequently occur. One common problem relates to monitoring tropical cyclones at night with only infrared-channel satellite data available. When strong vertical shear develops, often the convectively active portion of the storm can be separated from the lower portion of the storm circulation. At night, with just the infrared pictures, a shear-forced separation will not be detectable because the lower portion of the storm will be nearly invisible to the infrared sensors. When the first visible channel images are available in the morning and it becomes apparent that the storm has been sheared, there would have been overnight errors of current intensity by up to 15 m s^{-1} too high, position errors about 100 km or more, and several meters per second in storm motion error (Holland 1993). Real errors like this are similar to the type of contamination that we have built into the skill testing.

The results of this study suggest that large samples (i.e., $n \geq 100$) are needed for most forecasting studies of this type and that LAD regression is superior to LSD regression whenever a small amount of moderately contaminated data is present. The results also suggest that meteorological regression studies of the

Corresponding author address: Dr. Paul W. Mielke Jr., Department of Statistics, Colorado State University, Fort Collins, CO 80523.
E-mail: mielke@lamar.colostate.edu

type considered here should not be undertaken whenever a large amount of severely contaminated data is expected.

2. Description of statistical measures

Let the population and sample sizes be denoted by N and n, respectively; let y_i denote the dependent (predicted) variable; and let $x_{i1}, \ldots, x_{ip}$ denote the p independent (predictor) variables associated with the ith of n events. Consider the linear regression model given by

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i,$$

where $\beta_0, \ldots, \beta_p$ are $p + 1$ unknown parameters and e_i is an error term associated with the ith of n events. The LAD and LSD prediction equations are given by

$$\tilde{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} \tilde{\beta}_j x_{ij},$$

where $\tilde{y}_i$ is the predicted value of y_i and $\tilde{\beta}_0, \ldots, \tilde{\beta}_p$ minimize the expression

$$\sum_{i=1}^{n} |e_i|^v$$

and where $v = 1$ and $v = 2$ are associated with the LAD and LSD regression models, respectively.

The question arises as how best to determine the correspondence between the observed (y_i) values and the predicted ($\tilde{y}_i$) values, for $i = 1, \cdots, n$. A widely used method is to calculate the Pearson product-moment correlation coefficient (r) or the coefficient of determination (r^2) between the paired values (e.g., Barnston and Van den Dool 1993). The coefficient of determination is strictly a measure of linearity, and $r^2 = 1.0$ implies that all paired values of y_i and $\tilde{y}_i$ for $i = 1, \cdots, n$ fall on a line that does not necessarily have a unit slope nor passes though the origin. Consequently, an $r^2 = 1.0$ does not imply complete agreement between the paired y_i and $\tilde{y}_i$ values since $r^2 = 1.0$ if y_i and $\tilde{y}_i$ differ by an additive constant and/or by a multiplicative constant. For example, if the observed values are $y_i = i$ for $i = 1, \cdots, n$ and the corresponding predicted values are $\tilde{y}_i = 50 + 2i$ for $i = 1, \cdots, n$, then the coefficient of determination between y_i and $\tilde{y}_i$ is $r^2 = 1.0$; clearly, the prediction model that generated the $\tilde{y}_i$ values is useless. Thus, the use of the Pearson product-moment correlation coefficient in assessing prediction accuracy often produces an inflated index of forecast skill. To avoid this problem, either the mean square error (MSE) given by

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

or the root-mean-square error (rmse) given by

$$\text{rmse} = \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \right]^{1/2}$$

is often employed. While the MSE and rmse are zero when the y_i and $\tilde{y}_i$ values are identical, the two measures are not standardized measures and have no upper limits. In addition, neither measure is independent of the unit of measurement, and therefore, the measures are difficult to compare across studies. For example, if y_i is measured first in knots and second in meters per second, the values of the MSE and the rmse will change. Finally, both the MSE and rmse are conceptually misleading. The often-cited geometric representation of row vectors $(y_1, \ldots, y_n)$ and $(\tilde{y}_1, \ldots, \tilde{y}_n)$ in an n-dimensional space and the interpretation of $n^{1/2}$ rmse as the Euclidean distance between the observed and predicted n-dimensional points in this space is an artificial construct. In reality, the n paired values $\{(y_1, \tilde{y}_1), \ldots, (y_n, \tilde{y}_n)\}$ are n repeated pairs of points in a one-dimensional space. Furthermore, the MSE and the rmse involve squared Euclidean differences and they can be heavily influenced by one or more extreme values (Cotton et al. 1994), which are not uncommon in meteorological research. An alternative to the MSE or the rmse is the mean absolute error (MAE), in which the absolute differences are considered; that is,

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \tilde{y}_i|.$$

Like the MSE and the rmse, the MAE is not independent of the unit of measurement, is not a standardized measure, and has no upper limit. However, the MAE does mitigate the problem of extreme values. Finally, while the rmse is a minimum when the $\tilde{y}_i$ values are based on an LSD prediction model, the MAE is a minimum when the $\tilde{y}_i$ values are based on an LAD prediction model. Although the MAE is often computed on an LSD prediction model (e.g., Elsner and Schmertmann 1994), it is difficult to interpret when based on LSD regression, and when LSD regression is used, MAE values may not be comparable.

Because of the problems with these measures, many researchers have turned to measures of agreement in assessing prediction accuracy—for example, Willmott (1982), Willmott et al. (1985), Tucker et al. (1989), Gray et al. (1992), McCabe and Legates (1992), Badescu (1993), Elsner and Schmertmann (1993), Hess and Elsner (1994), and Cotton et al. (1994). For a recent comparison of various measures of agreement, see Watterson (1996).

In this study, the measure of agreement for both the LAD and LSD prediction equations is given by

$$\rho = 1 - \frac{\delta}{\mu_\delta},$$

where

$$\delta = \frac{1}{n}\sum_{i=1}^{n} |y_i - \tilde{y}_i|^{\upsilon},$$

$\upsilon = 1$ is associated with LAD regression, $\upsilon = 2$ is associated with LSD regression, and μ_δ is the average value of δ over all $n!$ equally likely permutations of $y_1, \ldots, y_n$ relative to $\tilde{y}_1, \ldots, \tilde{y}_n$ under the null hypothesis that the n pairs (y_i and $\tilde{y}_i$ for $i = 1, \cdots, n$) are merely the result of random assignment. This reduces to the simple computational form given by

$$\mu_\delta = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |y_i - \tilde{y}_j|^{\upsilon}.$$

Since ρ is a chance-corrected measure of agreement, $\rho = 1.0$ implies that all paired values of y_i and $\tilde{y}_i$ for $i = 1, \cdots, n$ fall on a line with unit slope that passes through the origin (i.e., a perfect forecast). Because δ = MAE when $\upsilon = 1$ and δ = MSE when $\upsilon = 2$, all values of ρ are based on $\upsilon = 1$ due to the geometric concern involving mse.

3. Construction of the population

In the notation of the previous section, the 3958 available primary events used to construct the five populations of this study consist of a dependent (predicted) variable (y) and 10 independent (predictor) variables ($x_1, \ldots, x_p$), where $p = 10$. The dependent ''variable'' for these populations is constructed from two datasets. In one dataset the predicted values are *intensity change* 24 h into the future; in the second dataset, the predicted values are *intensity* 24 h into the future. A simulation study of this type requires a population with a ρ value of approximately 0.50 in order to observe changes due to various sampling conditions and to reflect obtained ρ values in related studies (Gray et al. 1992, 1993, 1994). The descriptions for each of these variables follow:

- y Intensity change/intensity 24 h into the future (in knots)
- x_1 Julian date (e.g., 1 is 1 January and 365 is 31 December)
- x_2 Latitude (in degrees and tenths of degrees)
- x_3 Longitude (in degrees and tenths of degrees)
- x_4 Current intensity (in knots)
- x_5 Change of intensity in last 12 h (in knots)
- x_6 Change of intensity in last 24 h (in knots)
- x_7 Speed of storm in zonal direction (in knots, where positive is toward the east)
- x_8 Speed of storm in meridional direction (in knots, where positive is toward the north)
- x_9 Absolute magnitude of speed of storm (in knots)
- x_{10} Potential intensity difference (in knots) is based upon an exponential function of sea surface temperature (SST) minus the current intensity (DeMaria and Kaplan 1994)

The intensity change values and the 10 predictors were obtained from two separate datasets. Most of the values were constructed from the Atlantic basin best track data maintained by the National Hurricane Center (Jarvinen et al. 1984). Tropical storm and hurricane statistics of position, highest surface sustained winds, and lowest surface pressure (if measured) for every 6 h of their existence are available. Tropical storm data were removed for those storms that became extratropical or weakened below 35 kt by the 24-h verification time. Additionally, the SST data were obtained from the monthly SST (COADS) climatology (Reynolds 1988). These data are available on a $2° \times 2°$ grid based on the period 1950–79.

Two regression models designated as case 10 and case 6 are examined in this study. Case 10 involves all 10 independent variables ($p = 10$), whereas case 6 involves only 6 of the 10 independent variables (variables x_6, x_7, x_8, and x_9 are removed and $p = 6$). All datasets used in this study are available from the authors.

4. Simulation procedures

The present study investigates the effect of sample size, type of regression model (LAD and LSD), amount and degree of contamination, and noise-to-signal ratio on the degree of agreement between observed and predicted values in five populations that differ in amount and degree of contaminated data. Sample sizes (n) of 15, 25, 40, 65, 100, 160, 250, and 500 events are obtained from a fixed population of $N = 3958$ events that, for the purpose of this study, is not contaminated with extreme cases, a fixed population of $N = 3998$ events consisting of the initial population and 40 moderately extreme events (1% moderate contamination), a fixed population of $N = 3998$ events consisting of the initial population and 40 very extreme events (1% severe contamination), a fixed population of $N = 4158$ events consisting of the initial population and 200 moderately extreme events (5% moderate contamination), and a fixed population of $N = 4158$ events consisting of the initial population and 200 very extreme events (5% severe contamination).

The moderate 1% (5%) contamination consists of 40 (200) carefully designed additional events. The additional values of the independent variables were selected from the lowest and highest values of the specified independent variable in the initial population. Then, either the lowest or the highest value was selected, based on a random binary choice. The associated values of the dependent variable were selected from the center of the distribution of the dependent variable in the initial population, near the median. The severe 1% (5%) contamination involves 40 (200) centered dependent-variable values with the values of the independent variables placed at 2.5 times the lower and upper values of the ranges associated with the corre-

sponding independent variables in the initial population. The random sampling of events from each population was implemented in the bootstrap context; that is, the random sampling was accomplished with replacement. It should be noted that the contamination and examination of datasets containing extreme values is not new. Michaelsen (1987) analyzed datasets containing naturally occurring extreme values. Barnston and Van den Dool (1993) contaminated Gaussian datasets with extreme values in a study of cross-validated skill. As Barnston and Van den Dool (1993) note, extreme values are representative of many meteorological events and, in addition, inclusion of very extreme values, up to 10 standard deviations from the mean (Barnston and Van den Dool 1993), may be important as ''extreme design experiments.'' Finally, it should be emphasized that the initial population was designed and constructed from real data. The added events that contaminate the initial population create populations of data that are contaminated *relative* to the initial population. Whatever contamination preexists in the initial population of real data is unknown.

Two prediction models are considered for each population. The first prediction model (case 10) consists of $p = 10$ independent variables, and the second prediction model (case 6) consists of $p = 6$ independent variables. In case 10, 4 of the 10 independent variables were found to contribute no information to the predictions. Case 6 is merely the prediction model with the four noncontributing independent variables of case 10 deleted. Beginning with the 10 independent variables of case 10, backward selection was used on the entire population to identify the 6 contributing independent variables of case 6. The reason for the two prediction models is to examine the effect of including noncontributing independent variables in a prediction model.

A few caveats regarding the simulation study and its application to actual meteorological research follow. 1) Although the initial population is constructed from actual meteorological data, the purpose of this study is not to generate prediction models but rather to investigate statistical questions involving meteorological prediction methods. 2) While this study depends on the five specific populations that have been generated, the findings of this study are anticipated to hold for a variety of other populations. 3) Since this study involves random sampling from a fixed population, the results must be interpreted as a stationary process rather than an evolutionary process, which is commonly associated with climatic events in a time series framework. 4) In practice an investigator often exhausts the available data associated with a given study. Consequently, the results of the present study should be interpreted in the context of an unknown population for which a prediction model is desired. 5) The present study is strictly an empirical study. No distributional assumptions (e.g., normality) are made about any of the variables in the population. 6) The purpose of this study is to help investigators choose sample sizes and regression techniques for future research. Because of the artificial nature of the dependent variable, no predictions to actual meteorological events are intended.

5. Discussion of the findings

The findings of the study are summarized in Tables 1–5. In Tables 1a, 2a, 3a, 4a, and 5a, each row is specified by 1) a sample size (n), 2) $p = 10$ (case 10) and $p = 6$ (case 6) independent variables, and 3) LAD and LSD regression analyses. In each of these tables the first column (C1) contains the true ρ values for the designated population, the second column (C2) contains the average of 10 000 randomly obtained sample $\hat{\rho}$ values of a specified size where the $\tilde{y}$ values are based on the true population regression coefficients, and the third column (C3) contains the average of 10 000 randomly obtained sample $\hat{\rho}$ values where the $\tilde{y}$ values are based on the sample regression coefficients for each of the 10 000 independent samples. The fourth column (C4) is more complicated and is designed to measure the effectiveness of validating sample regression coefficients. Here the sample regression coefficients from 10 000 random samples are obtained from column C3; then, for each of these 10 000 sets of sample regression coefficients an additional set of five independent random samples of the same designated size ($n = 15, \cdots, 500$) are drawn from the population. The sample regression coefficients from C3 are then applied to each of these five new samples, and $\hat{\rho}$ values are computed for each of these five samples for a total of 50 000 $\hat{\rho}$ values. The average of these 50 000 $\hat{\rho}$ values is reported in column C4 of Table 1a, yielding a measure of the effectiveness of sample validation—that is, applying the sample regression coefficients from a single sample to five new independent samples drawn from the same population.

In Tables 1b, 2b, 3b, 4b, and 5b, each row is specified by a sample size (n), $p = 10$ (case 10) and $p = 6$ (case 6) independent variables, and LAD and LSD regression analyses. In each of these tables the first column (C2/C1) contains the ratio of the average $\hat{\rho}$ value of C2 to the corresponding true population ρ value of C1, the second column (C3/C1) contains the ratio of the average $\hat{\rho}$ value of C3 to the corresponding true population ρ value of C1, the third column (C4/C1) contains the ratio of the average $\hat{\rho}$ value of C4 to the corresponding true population ρ value of C1, and the fourth column (C4/C3) contains the ratio of the average $\hat{\rho}$ value of C4 to the average $\hat{\rho}$ value of C3.

a. Overview of the findings

There are four types of predictive skill to be examined in this study: true skill, optimal skill, artificial skill, and expected skill. Each of the four types is considered under the following conditions: contamination of the

TABLE 1a. Population 1: Initial population consisting of 3958 noncontaminated events. Columns are (C1) true population ρ values, (C2) average of 10 000 sample $\hat{\rho}$ values based on population regression coefficients, (C3) average of 10 000 sample $\hat{\rho}$ values based on regression coefficients for each sample, and (C4) average of 5 sample $\hat{\rho}$ values for each of 10 000 random sets of regression coefficients associated with the 10 000 samples of C3.

Sample size	Case	Model	C1	C2	C3	C4
15	10	LAD	0.51495	0.48082	0.83216	0.21959
		LSD	0.51154	0.47761	0.76579	0.24721
	6	LAD	0.51130	0.47737	0.69947	0.32214
		LSD	0.50917	0.47552	0.64059	0.34883
25	10	LAD	0.51495	0.49471	0.69659	0.34693
		LSD	0.51154	0.49153	0.63931	0.37427
	6	LAD	0.51130	0.49102	0.61963	0.39741
		LSD	0.50917	0.48911	0.57839	0.41795
40	10	LAD	0.51495	0.50259	0.62613	0.41279
		LSD	0.51154	0.49937	0.58533	0.43132
	6	LAD	0.51130	0.49918	0.57687	0.43965
		LSD	0.50917	0.49717	0.54955	0.45455
65	10	LAD	0.51495	0.50770	0.58265	0.45361
		LSD	0.51154	0.50438	0.55438	0.46425
	6	LAD	0.51130	0.50414	0.55102	0.46701
		LSD	0.50917	0.50212	0.53274	0.47611
100	10	LAD	0.51495	0.51068	0.55843	0.47627
		LSD	0.51154	0.50729	0.53790	0.48182
	6	LAD	0.51130	0.50702	0.53651	0.48269
		LSD	0.50917	0.50499	0.52353	0.48814
160	10	LAD	0.51495	0.51224	0.54290	0.49184
		LSD	0.51154	0.50890	0.52759	0.49302
	6	LAD	0.51130	0.50855	0.52727	0.49364
		LSD	0.50917	0.50650	0.51780	0.49598
250	10	LAD	0.51495	0.51352	0.53325	0.50076
		LSD	0.51154	0.51010	0.52141	0.49982
	6	LAD	0.51130	0.50987	0.52160	0.50012
		LSD	0.50917	0.50771	0.51454	0.50081
500	10	LAD	0.51495	0.51363	0.52527	0.50865
		LSD	0.51154	0.51017	0.51661	0.50562
	6	LAD	0.51130	0.50992	0.51685	0.50578
		LSD	0.50917	0.50779	0.51206	0.50500

population data in both degree and amount; type of regression model used, that is, LAD and LSD; the ratio of noise-to-signal in the data where the 10-predictor model (case 10) contains a relatively high noise-to-signal ratio and the 6-predictor model (case 6) contains a relatively low noise-to-signal ratio; and sample size, which varies from $n = 15$ to $n = 500$.

The first type of skill to be considered is *true skill*, which is defined as the agreement, measured by a ρ value, between the observed (y) and predicted ($\tilde{y}$) values when the entire population is available and the ρ values are based on the true population regression coefficients. In general, true skill is used as a benchmark against which the other three forms of skill are evaluated. The true skill ρ values are given in column C1 of Tables 1a, 2a, 3a, 4a, and 5a.

The second type of skill is *optimal skill*, which reflects the average agreement, measured by a $\hat{\rho}$ value, between the observed (y) and predicted ($\tilde{y}$) values when only a specified sample is available and the $\tilde{y}$ values are based on the true population regression coefficients. Optimal skill is measured as the ratio of the sample $\hat{\rho}$ value, with the population regression coefficients presumed known, to the corresponding true skill ρ value. Specifically, the relevant $\hat{\rho}$ values are given in column C2 of Tables 1a, 2a, 3a, 4a, and 5a, and the optimal skill ratios are given in the C2/C1 column of Tables 1b, 2b, 3b, 4b, and 5b. The expectation is that the tabled C2/C1 ratios will be a little less than 1.0, even for small samples, because they reflect what would happen if a researcher drew a sample and, fortuitously, happened to get a set of sample-based regression coefficients very close to the true population regression coefficients. The reason that the tabled C2/C1 ratios are not equal to 1.0 is because the sum of errors in the sample is not minimized by the population regression coefficients. Note, however, that as sample size increases the corresponding sample optimal fit approaches the population fit and the C2/C1 values approach 1.0.

The third type of skill is *artificial skill*, which reflects the average agreement, measured by a $\hat{\rho}$ value, between the observed (y) and predicted ($\tilde{y}$) values when a specified sample is available and the $\tilde{y}$ values are based on the sample regression coefficients (Shapiro 1984). For an alternative definition of artificial skill, based on the

TABLE 1b. Population 1: Initial population consisting of 3958 noncontaminated events. Columns are ratio estimators C2/C1, C3/C1, C4/C1, and C4/C3 associated with C1, C2, C3, and C4 of Table 1a.

Sample size	Case	Model	C2/C1	C3/C1	C4/C1	C4/C3
15	10	LAD	0.934	1.616	0.426	0.264
		LSD	0.934	1.497	0.483	0.323
	6	LAD	0.934	1.368	0.630	0.461
		LSD	0.934	1.258	0.685	0.545
25	10	LAD	0.961	1.353	0.674	0.498
		LSD	0.961	1.250	0.732	0.585
	6	LAD	0.961	1.212	0.777	0.641
		LSD	0.961	1.136	0.821	0.723
40	10	LAD	0.976	1.216	0.802	0.659
		LSD	0.976	1.144	0.843	0.737
	6	LAD	0.976	1.128	0.860	0.762
		LSD	0.976	1.080	0.893	0.827
65	10	LAD	0.986	1.131	0.881	0.779
		LSD	0.986	1.084	0.908	0.837
	6	LAD	0.986	1.078	0.913	0.848
		LSD	0.986	1.046	0.935	0.894
100	10	LAD	0.992	1.084	0.925	0.853
		LSD	0.992	1.052	0.942	0.896
	6	LAD	0.992	1.049	0.944	0.900
		LSD	0.992	1.028	0.959	0.932
160	10	LAD	0.995	1.054	0.955	0.906
		LSD	0.995	1.031	0.965	0.934
	6	LAD	0.995	1.031	0.965	0.936
		LSD	0.995	1.017	0.974	0.958
250	10	LAD	0.997	1.036	0.972	0.939
		LSD	0.997	1.019	0.977	0.959
	6	LAD	0.997	1.020	0.978	0.959
		LSD	0.997	1.011	0.984	0.973
500	10	LAD	0.997	1.013	0.988	0.968
		LSD	0.997	1.010	0.988	0.985
	6	LAD	0.997	1.011	0.989	0.979
		LSD	0.997	1.006	0.992	0.986

difference between hindcast and forecast skill, see Michaelsen (1987). Artificial skill is measured as the ratio of the sample $\hat{\rho}$ value, with the population regression coefficients presumed unknown, to the corresponding true skill ρ value. Specifically, the relevant $\hat{\rho}$ values are given in column C3 of Tables 1a, 2a, 3a, 4a, and 5a, and the artificial skill ratios are given in the C3/C1 column of Tables 1b, 2b, 3b, 4b, and 5b. The expectation is that the C3/C1 values will be slightly above 1.0 because of what is commonly termed ''retrospective'' fit (Copas 1983) between the y and $\tilde{y}$ values; that is, the sum of errors is minimized because the regression coefficients are based on the sample data. In general, tabled C3/C1 values greater than 1.0 reflect the amount of artificial skill inherent in retrospective fit; for convenience, we will call this a ''degrading'' of the prediction; that is, the sample-based $\hat{\rho}$ value overestimates the true population ρ value and the sample $\hat{\rho}$ value must be degraded by multiplying it by the reciprocal of the tabled C3/C1 value.

It should be noted in this context that artificial skill is a type of optimizing bias where the results are biased upward. There is a second type of bias that also contributes to artificial skill: selection bias. This occurs when a subset of independent variables is selected from the population based on information in the sample. As in the case of optimizing bias, artificial skill is biased upward when selection bias is present. In this study, the measure of artificial skill reflects only optimizing bias. Selection bias has been controlled by selecting the two sets of independent variables (cases 10 and 6) from information contained in the designed population, not from information contained in a sample.

The fourth type of skill is *expected skill,* which reflects the average agreement, measured by a $\hat{\rho}$ value, between the observed (y) and predicted ($\tilde{y}$) values when a specified sample is available and the sample regression coefficients are applied to an additional set of samples independently drawn from the same population. Expected skill is measured as the ratio of the average sample $\hat{\rho}$ value to the corresponding true skill ρ value. The relevant $\hat{\rho}$ values are given in column C4 of Tables 1a–5a, and the expected skill ratios are given in the C4/C1 column of Tables 1b–5b. The expectation is that the tabled C4/C1 values will be slightly less than 1.0 because they reflect what is commonly termed ''prospective'' or ''validation'' fit (Copas 1983) between the y and $\tilde{y}$ values; that is, the sum of errors is not minimized because the regression coefficients are based on only one of the six independently drawn ran-

TABLE 2a. Population 2: Contaminated population of 3998 events consisting of the initial population of 3958 events and 40 moderately extreme events. Columns are (C1) true population ρ values, (C2) average of 10 000 sample $\hat{\rho}$ values based on population regression coefficients, (C3) average of 10 000 sample $\hat{\rho}$ values based on regression coefficients for each sample, and (C4) average of 5 sample $\hat{\rho}$ values for each of 10 000 random sets of regression coefficients associated with the 10 000 samples of C3.

Sample size	Case	Model	C1	C2	C3	C4
15	10	LAD	0.48886	0.45910	0.83077	0.20662
		LSD	0.45120	0.42411	0.76387	0.23315
	6	LAD	0.48220	0.45252	0.69081	0.30215
		LSD	0.44984	0.42341	0.63008	0.32887
25	10	LAD	0.48886	0.46911	0.69099	0.32943
		LSD	0.45120	0.43261	0.63311	0.35703
	6	LAD	0.48220	0.46237	0.60467	0.37074
		LSD	0.44984	0.43126	0.56220	0.39249
40	10	LAD	0.48886	0.47813	0.61657	0.38947
		LSD	0.45120	0.44118	0.57548	0.40904
	6	LAD	0.48220	0.47150	0.55632	0.40805
		LSD	0.44984	0.43999	0.52658	0.42208
65	10	LAD	0.48886	0.48137	0.56622	0.42434
		LSD	0.45120	0.44442	0.53715	0.43587
	6	LAD	0.48220	0.47475	0.52586	0.43418
		LSD	0.44984	0.44322	0.50036	0.43615
100	10	LAD	0.48886	0.48438	0.53914	0.44413
		LSD	0.45120	0.44706	0.51555	0.44819
	6	LAD	0.48220	0.47785	0.51103	0.45081
		LSD	0.44984	0.44588	0.48556	0.44265
160	10	LAD	0.48886	0.48586	0.51867	0.45922
		LSD	0.45120	0.44862	0.49590	0.45266
	6	LAD	0.48220	0.47922	0.49982	0.46294
		LSD	0.44984	0.44729	0.47246	0.44554
250	10	LAD	0.48886	0.48792	0.50767	0.46914
		LSD	0.45120	0.45021	0.48257	0.45410
	6	LAD	0.48220	0.48128	0.49408	0.47041
		LSD	0.44984	0.44889	0.46486	0.44727
500	10	LAD	0.48886	0.48793	0.49896	0.47922
		LSD	0.45120	0.45036	0.46904	0.45367
	6	LAD	0.48220	0.48131	0.48999	0.47780
		LSD	0.44984	0.44905	0.45838	0.44897

dom samples. In general, tabled C4/C1 values less than 1.0 indicate the amount of skill that is expected relative to the true skill possible when a population is available. More specifically, if researchers were to use the sample coefficients in a prediction equation, as is commonly done in practice, then the C4/C1 values indicate the expected reduction in fit of the y and $\tilde{y}$ values for future results. Any tabled C4/C1 value greater than 1.0 is cause for concern since this indicates that the sample estimates of the population regression coefficients provide a better validation fit, on average, than would be possible had the actual population been available and is evidence that some sort of inflation of expected skill is present in the analysis.

b. Population 1

Population 1 is the initial population of $N = 3958$ noncontaminated events. The results of the analysis of population 1 are summarized in Tables 1a and 1b. Since sample $\hat{\rho}$ values that are based on true population regression coefficients behave very much like unbiased estimators of the true ρ values, the average sample $\hat{\rho}$ values in column C2 of Table 1a are, as expected, very close to the true population ρ values in column C1 of Table 1a. The corresponding ratios are given in the C2/C1 column of Table 1b. It is obvious from an inspection of these values that larger sample sizes provide better predictions; that is, the ratio approaches 1.0 as sample size increases from $n = 15$ to $n = 500$, there are no differences between the 10-predictor model (case 10) and the 6-predictor model (case 6), and there are no appreciable differences between the LAD and LSD regression models, other conditions being equal.

In most studies, the population regression coefficients are not known, and the sample $\hat{\rho}$ value is based strictly on the sample regression coefficients, as in column C3 of Table 1a. Whenever a sample is obtained from a population there will be, on average, a degrading of the prediction; that is, the sample-based $\hat{\rho}$ value will overestimate the true population ρ value. Column C3/C1 in Table 1b is a measure of the degrading for this population. Inspection of the C3/C1 column indicates that the sample $\hat{\rho}$ values are indeed biased upward, as all of the C3/C1 values are greater than 1.0.

TABLE 2b. Population 2: Contaminated population of 3998 events consisting of the initial population of 3958 events and 40 moderately extreme events. Columns are ratio estimators C2/C1, C3/C1, C4/C1, and C4/C3 associated with C1, C2, C3, and C4 of Table 2a.

Sample size	Case	Model	C2/C1	C3/C1	C4/C1	C4/C3
15	10	LAD	0.937	1.699	0.423	0.249
		LSD	0.940	1.693	0.517	0.305
	6	LAD	0.938	1.433	0.627	0.437
		LSD	0.941	1.401	0.731	0.522
25	10	LAD	0.960	1.413	0.674	0.477
		LSD	0.959	1.403	0.791	0.564
	6	LAD	0.959	1.254	0.769	0.613
		LSD	0.959	1.250	0.873	0.698
40	10	LAD	0.978	1.261	0.797	0.632
		LSD	0.978	1.275	0.907	0.711
	6	LAD	0.978	1.154	0.846	0.733
		LSD	0.978	1.171	0.938	0.802
65	10	LAD	0.985	1.158	0.868	0.749
		LSD	0.985	1.190	0.966	0.811
	6	LAD	0.985	1.091	0.900	0.826
		LSD	0.985	1.112	0.970	0.872
100	10	LAD	0.991	1.103	0.909	0.824
		LSD	0.991	1.143	0.993	0.869
	6	LAD	0.991	1.060	0.935	0.882
		LSD	0.991	1.079	0.984	0.912
160	10	LAD	0.994	1.061	0.939	0.885
		LSD	0.994	1.099	1.003	0.913
	6	LAD	0.994	1.037	0.960	0.926
		LSD	0.994	1.050	0.990	0.943
250	10	LAD	0.998	1.038	0.960	0.924
		LSD	0.998	1.070	1.006	0.941
	6	LAD	0.998	1.025	0.964	0.952
		LSD	0.998	1.033	0.994	0.962
500	10	LAD	0.998	1.021	0.980	0.960
		LSD	0.998	1.040	1.005	0.967
	6	LAD	0.998	1.016	0.991	0.975
		LSD	0.998	1.019	0.998	0.979

It is also clear that the amount of degrading decreases with increasing sample size, that case 6 (6 predictors) is superior to case 10 (10 predictors), and that the LSD regression model provides less bias than the LAD regression model. Note also that most of these differences disappear as the sample size becomes larger.

A comparison of the C2/C1 and C3/C1 columns yields an important conclusion: should an investigator be fortunate enough to select a sample, of any size, that yields regression coefficients that are close to the true population regression coefficients, then, as can be seen in column C2/C1, the predicted values will show high agreement with the observed values. A luxury of a simulation study of this type is that the true population values are known. In most research situations, an investigator has only a single sample with which to work and has no way of knowing if the obtained $\hat{\rho}$ value is too high.

The problem of validated predictions (really predictions which are not validated) is important in meteorological forecasting, and the C4/C1 column provides a measure of the effectiveness of validated predictions for population 1. The values presented in the C4/C1 column indicate that validation is extremely poor for small sample sizes where the expected skill ratios are considerably less than 1.0, but the problem nearly disappears for larger samples. There is a considerable difference between case 10 and case 6 for small samples, but most of the difference disappears for the larger samples. The LSD regression model is superior to the LAD regression model for small samples, but there is no difference for the larger samples. Note that no tabled C4/C1 value exceeds 1.0. Values greater than 1.0 would indicate, as noted previously, that sample estimates of the population regression coefficients provide better validation fits, on average, than would be possible had the actual population been available. The ratio values in the C4/C3 column contain the amount of expected skill (C4/C1) adjusted for the amount of artificial skill (C3/C1). There is an amount by which validation fit (C4) falls short of retrospective fit (C3), and the values in the C4/C3 column summarize this ''shrinkage'' (cf. Copas 1983) in a ratio format. The C4/C3 ratio provides the most stringent measure of anticipated reduction in prediction, relative to available sample information. While it is not possible to compare across cases (10 and 6) or across models (LAD and LSD) because a common base does not exist, it is possible to compare across sample size ($n = 15$ to $n = 500$) within the same case and model. It is clear that, within these restrictions,

TABLE 3a. Population 3: Contaminated population of 3998 events consisting of the initial population of 3958 events and 40 very extreme events. Columns are (C1) true population ρ values, (C2) average of 10 000 sample $\hat{\rho}$ values based on population regression coefficients, (C3) average of 10 000 sample $\hat{\rho}$ values based on regression coefficients for each sample, and (C4) average of 5 sample $\hat{\rho}$ values for each of 10 000 random sets of regression coefficients associated with the 10 000 samples of C3.

Sample size	Case	Model	C1	C2	C3	C4
15	10	LAD	0.44873	0.43134	0.83121	0.20082
		LSD	0.29776	0.28300	0.76468	0.22665
	6	LAD	0.43722	0.42140	0.69002	0.29357
		LSD	0.27225	0.26012	0.62930	0.31866
25	10	LAD	0.44873	0.43807	0.69172	0.31845
		LSD	0.29776	0.28593	0.63445	0.34486
	6	LAD	0.43722	0.42716	0.60065	0.35366
		LSD	0.27225	0.26222	0.55827	0.37435
40	10	LAD	0.44873	0.44390	0.61802	0.37532
		LSD	0.29776	0.29164	0.57769	0.39366
	6	LAD	0.43722	0.43308	0.54451	0.37879
		LSD	0.27225	0.26747	0.51698	0.39520
65	10	LAD	0.44873	0.44517	0.56667	0.40517
		LSD	0.29776	0.29398	0.53918	0.41701
	6	LAD	0.43722	0.43396	0.49965	0.38841
		LSD	0.27225	0.26908	0.47894	0.39580
100	10	LAD	0.44873	0.44689	0.53523	0.41757
		LSD	0.29776	0.29530	0.51541	0.42518
	6	LAD	0.43722	0.43560	0.47492	0.39691
		LSD	0.27225	0.27015	0.44794	0.38686
160	10	LAD	0.44873	0.44739	0.50458	0.42088
		LSD	0.29776	0.29634	0.48765	0.42113
	6	LAD	0.43722	0.43595	0.45708	0.40649
		LSD	0.27225	0.27109	0.41056	0.36825
250	10	LAD	0.44873	0.44871	0.48421	0.42473
		LSD	0.29776	0.29699	0.45964	0.40951
	6	LAD	0.43722	0.43730	0.44837	0.41433
		LSD	0.27225	0.27167	0.37856	0.34883
500	10	LAD	0.44873	0.44838	0.46656	0.43340
		LSD	0.29776	0.29723	0.40724	0.37716
	6	LAD	0.43722	0.43696	0.44242	0.42437
		LSD	0.27225	0.27190	0.33445	0.31859

larger sample sizes yield less shrinkage of expected skill than smaller sample sizes. For example, given case 6, the LAD regression model, and $n = 15$, C4/C3 = 0.461; however, C4/C3 = 0.762 when $n = 40$. Shrinkage in expected skill is minimal for $n \geqslant 250$.

Finally, with respect to Table 1a, a feature worth noting is that a true population ρ value of 1.0 implies that the corresponding $\hat{\rho}$ values in columns C2, C3, and C4 must also be 1.0. A population value of $\rho = 1.0$ reflects a perfect linear relationship between the observed and predicted values; that is, all points are on a straight line having unit slope and zero origin. Consequently, any sample of values selected from such a population must necessarily yield a $\hat{\rho}$ value of 1.0.

c. Population 2

Population 2 is the contaminated population of $N = 3998$ events consisting of the initial population of 3958 events and 40 moderately extreme events (i.e., 1% moderate contamination). The results of the analysis of population 2 are summarized in Tables 2a and 2b. The average sample $\hat{\rho}$ values in column C2 of Table 2a are very close to the true population ρ values in column C1, and as can be seen in the C2/C1 column of Table 2b, larger sample sizes provide better predictions than smaller samples. In addition, there is no difference between the LAD and LSD regression models, and very little difference between case 10 and case 6. It should be noted that case 10 and case 6 are well defined only for the initial population (population 1), which contains no contaminated data. That is, in the initial uncontaminated population, the four additional predictors of case 10 are truly noise and add no predictive power above and beyond the 6 predictors of case 6. However, in populations 2–5 the difference between case 10 and case 6 is not necessarily only noise. Because the contamination has been added to the independent variables, it may be that for populations 2–5 the four additional predictors of case 10 now contain a real signal.

Examination of the C3/C1 column in Table 2b reveals that degrading follows the same general pattern as the C3/C1 column of Table 1b except that there is more degrading due to the contamination of the population. Again, the amount of degrading decreases with

TABLE 3b. Population 3: Contaminated population of 3998 events consisting of the initial population of 3958 events and 40 very extreme events. Columns are ratio estimators C2/C1, C3/C1, C4/C1, and C4/C3 associated with C1, C2, C3, and C4 of Table 3a.

Sample size	Case	Model	C2/C1	C3/C1	C4/C1	C4/C3
15	10	LAD	0.961	1.852	0.448	0.242
		LSD	0.950	2.568	0.761	0.296
	6	LAD	0.964	1.578	0.671	0.425
		LSD	0.955	2.311	1.170	0.506
25	10	LAD	0.976	1.538	0.710	0.460
		LSD	0.960	2.131	1.158	0.544
	6	LAD	0.977	1.374	0.809	0.589
		LSD	0.963	2.051	1.375	0.671
40	10	LAD	0.989	1.337	0.836	0.607
		LSD	0.979	1.940	1.322	0.681
	6	LAD	0.991	1.245	0.866	0.696
		LSD	0.982	1.899	1.452	0.764
65	10	LAD	0.992	1.263	0.903	0.715
		LSD	0.987	1.811	1.400	0.773
	6	LAD	0.993	1.143	0.888	0.777
		LSD	0.988	1.759	1.454	0.826
100	10	LAD	0.996	1.193	0.931	0.780
		LSD	0.992	1.731	1.428	0.825
	6	LAD	0.996	1.086	0.908	0.836
		LSD	0.992	1.645	1.421	0.864
160	10	LAD	0.997	1.124	0.938	0.834
		LSD	0.995	1.638	1.414	0.864
	6	LAD	0.997	1.045	0.930	0.889
		LSD	0.996	1.508	1.353	0.897
250	10	LAD	1.000	1.079	0.947	0.877
		LSD	0.997	1.544	1.375	0.891
	6	LAD	1.000	1.026	0.948	0.924
		LSD	0.998	1.390	1.281	0.921
500	10	LAD	0.999	1.040	0.966	0.929
		LSD	0.998	1.368	1.267	0.926
	6	LAD	0.999	1.012	0.971	0.959
		LSD	0.999	1.228	1.170	0.953

increasing sample size, and case 6 is consistently superior to case 10. However, in contrast to Table 1b, it is the LAD regression model that provides less bias than the LSD regression model once sample size is increased to about $n = 40$. Column C4/C1 of Table 2b reflects the same general pattern of expected skill as column C4/C1 of Table 1b except that the validation fit is worse overall due to the contamination of the population. Again, the validation fit is poor for small samples. Case 6 generally does better than case 10, and the LSD regression model performs better than the LAD regression model, for most cases. One problematic result appears in the C4/C1 column for case 10, LSD, and $n = 160$, $n = 250$, and $n = 500$, where the average $\hat{\rho}$ values exceed 1.0. In all three cases this occurs with large n, with the LSD regression model, and with case 10, which contains the four predictors that add nothing to the prediction equation in population 1. These results are consistent with the findings of Barnston and Van den Dool (1993) in their study of cross-validation skill.

As noted previously, the LSD regression model appears to do better than the LAD regression model for small to moderate sample sizes. However, for larger sample sizes the LSD regression model is clearly biased, overstating the validation fit and producing exaggerated estimates of expected skill. It appears, from the larger sample results, that LSD may not really be doing better than LAD for smaller sample sizes and that exaggerated skill may also be present in these small sample results without pushing the ratios over 1.0. The tabled C4/C1 values greater than 1.0 indicate that the LSD regression model provides a validation fit that is too optimistic and casts a shadow of suspicion on those LSD regression results that are higher than the LAD regression results but still less than 1.0. The C4/C3 values in Table 2b indicate that the shrinkage of expected skill decreases as sample size increases, and shrinkage is of little consequence for $n \geqslant 250$.

d. Population 3

Population 3 is the contaminated population of $N = 3998$ events consisting of the initial population of 3958 events and 40 very extreme events (i.e., 1% severe contamination). The results of the analysis of population 3 are summarized in Tables 3a and 3b. Inspection of Table 3b discloses that even a small amount (i.e., 1%) of severe contamination of a population produces acute problems for the LSD regression model. In this population there is a small amount of severe con-

TABLE 4a. Population 4: Contaminated population of 4158 events consisting of the initial population of 3958 events and 200 moderately extreme events. Columns are (C1) true population ρ values, (C2) average of 10 000 sample $\hat{\rho}$ values based on population regression coefficients, (C3) average of 10 000 sample $\hat{\rho}$ values based on regression coefficients for each sample, and (C4) average of 5 sample $\hat{\rho}$ values for each of 10 000 random sets of regression coefficients associated with the 10 000 samples of C3.

Sample size	Case	Model	C1	C2	C3	C4
15	10	LAD	0.36924	0.34738	0.82319	0.17630
		LSD	0.31192	0.29274	0.75362	0.19886
	6	LAD	0.36698	0.34557	0.66307	0.24413
		LSD	0.30599	0.28720	0.59807	0.26934
25	10	LAD	0.36924	0.35603	0.66658	0.26978
		LSD	0.31192	0.30014	0.60568	0.29496
	6	LAD	0.36698	0.35403	0.54840	0.28455
		LSD	0.30599	0.29463	0.50205	0.30704
40	10	LAD	0.36924	0.36334	0.57417	0.30939
		LSD	0.31192	0.30653	0.52983	0.33079
	6	LAD	0.36698	0.36115	0.48241	0.30925
		LSD	0.30599	0.30071	0.44501	0.32037
65	10	LAD	0.36924	0.36553	0.50310	0.33020
		LSD	0.31192	0.30853	0.46771	0.34046
	6	LAD	0.36698	0.36338	0.43752	0.32591
		LSD	0.30599	0.30268	0.39896	0.32084
100	10	LAD	0.36924	0.36664	0.45960	0.34149
		LSD	0.31192	0.30930	0.42314	0.33801
	6	LAD	0.36698	0.36439	0.41285	0.33750
		LSD	0.30599	0.30348	0.36963	0.31813
160	10	LAD	0.36924	0.36788	0.42656	0.34929
		LSD	0.31192	0.31047	0.38503	0.33172
	6	LAD	0.36698	0.36563	0.39421	0.34556
		LSD	0.30599	0.30464	0.34668	0.31471
250	10	LAD	0.36924	0.36784	0.40505	0.35441
		LSD	0.31192	0.31052	0.35909	0.32608
	6	LAD	0.36698	0.36567	0.38328	0.35161
		LSD	0.30599	0.30475	0.33194	0.31188
500	10	LAD	0.36924	0.36859	0.38708	0.36071
		LSD	0.31192	0.31133	0.33582	0.32013
	6	LAD	0.36698	0.36635	0.37622	0.36011
		LSD	0.30599	0.30544	0.31932	0.30961

tamination, and an examination of the C2/C1 column in Table 3b shows that larger sample sizes increase accuracy, that case 6 is superior to case 10, and that, for the first time, the LAD regression model performs better than the LSD regression model. Column C3/C1 underscores the problem of estimating population parameters from sample statistics whenever contamination is present. Obviously, there is a large amount of artificial skill, indicated by values considerably greater than 1.0. Here, as in the C2/C1 column, case 6 is better than case 10, and the LAD regression model consistently outperforms the LSD regression model. Finally, the results listed in the C4/C1 column reveal acute problems with the LSD regression model, which introduces severe inflation of expected skill for both case 10 and case 6 and for nearly all sample sizes. On the other hand, the LAD regression model does very well, provided the sample size is greater than $n = 65$. The C4/C3 ratio values in Table 3b are similar to those in Tables 1b and 2b: there is little shrinkage for the larger samples, and shrinkage decreases as sample size increases.

e. Population 4

Population 4 is the contaminated population of $N = 4158$ events consisting of the initial population of 3958 events and 200 moderately extreme events (i.e., 5% moderate contamination). The results of the analysis of population 4 are summarized in Tables 4a and 4b. Inspection of column C2/C1 of Table 4b yields the following conclusions: the larger sample sizes yield better results except for the largest sample sizes in which the results are approximately equal, case 6 performs better than case 10, and the LAD regression model is much better than the LSD regression model. Column C3/C1 clearly shows the bias in estimation when relying on sample estimators. While very large samples control for this bias, results with smaller sample sizes are clearly biased upward with much artificial skill in evidence. Again, case 6 performs better than case 10, and the LAD regression model is consistently superior to the LSD regression model. Column C4/C1 clearly demonstrates the inherent difficulties with the LSD regression model, where almost all of the ratios

TABLE 4b. Population 4: Contaminated population of 4158 events consisting of the initial population of 3958 events and 200 moderately extreme events. Columns are ratio estimators C2/C1, C3/C1, C4/C1, and C4/C3 associated with C1, C2, C3, and C4 of Table 4a.

Sample size	Case	Model	C2/C1	C3/C1	C4/C1	C4/C3
15	10	LAD	0.941	2.229	0.477	0.214
		LSD	0.939	2.416	0.638	0.242
	6	LAD	0.942	1.807	0.665	0.368
		LSD	0.939	1.955	0.880	0.450
25	10	LAD	0.964	1.805	0.731	0.405
		LSD	0.962	1.942	0.946	0.487
	6	LAD	0.965	1.494	0.775	0.519
		LSD	0.963	1.641	1.003	0.612
40	10	LAD	0.984	1.555	0.838	0.539
		LSD	0.983	1.699	1.060	0.624
	6	LAD	0.984	1.315	0.843	0.641
		LSD	0.983	1.454	1.047	0.720
65	10	LAD	0.990	1.363	0.894	0.656
		LSD	0.989	1.499	1.091	0.728
	6	LAD	0.990	1.192	0.888	0.745
		LSD	0.989	1.304	1.049	0.804
100	10	LAD	0.993	1.245	0.925	0.743
		LSD	0.992	1.357	1.084	0.799
	6	LAD	0.993	1.125	0.920	0.817
		LSD	0.992	1.208	1.040	0.861
160	10	LAD	0.996	1.155	0.946	0.819
		LSD	0.995	1.234	1.063	0.862
	6	LAD	0.996	1.074	0.942	0.877
		LSD	0.996	1.133	1.028	0.908
250	10	LAD	0.996	1.097	0.960	0.875
		LSD	0.996	1.151	1.045	0.908
	6	LAD	0.996	1.044	0.958	0.917
		LSD	0.996	1.085	1.019	0.940
500	10	LAD	0.998	1.048	0.977	0.932
		LSD	0.998	1.077	1.026	0.953
	6	LAD	0.998	1.025	0.981	0.957
		LSD	0.998	1.044	1.012	0.970

exceed 1.0. Here, the LAD regression model provides excellent validation fits and reasonable estimates of expected skill, while the LSD regression model is clearly overfitting the y and $\tilde{y}$ values and providing inflated estimates of expected skill. The C4/C3 ratio values in Table 4b indicate that as sample size increases, shrinkage decreases.

f. Population 5

Population 5 is the contaminated population of $N = 4158$ events consisting of the initial population of 3958 events and 200 very extreme events (i.e., 5% severe contamination). The results of the analysis of population 5 are summarized in Tables 5a and 5b. The C2/C1 values are not that different from the C2/C1 values of the other populations. The values in column C3/C1 show just how bad sample estimators can be, especially for small samples, with many estimates of artificial skill in excess of 4.0 and some even greater than 5.0. It is interesting to note, however, that the LAD regression model performs better than the LSD regression model for all sample sizes. In addition, case 10 outperforms case 6 in these circumstances. Column C4/C1 demonstrates the problem of generalizing to other samples with contaminated population data. Here, both the LAD and LSD regression models incorporate inflated expected skill, although the LAD regression model is less affected than the LSD regression model. As in Tables 1b, 2b, 3b, and 4b, the C4/C3 values indicate that as sample size is increased, shrinkage decreases. A comparison of the C4/C3 values across Tables 1b, 2b, 3b, 4b, and 5b reveals that as contamination is increased in amount and severity, the C4/C3 shrinkage ratio values decrease. However, the decrease in the C4/C3 ratio values for Tables 1b, 2b, 3b, 4b, and 5b is quite small relative to the amount of contamination introduced.

6. The problem of degrading

a. General discussion

An obvious common feature of the results in Tables 1–5 is the degrading of the sample $\hat{\rho}$ values that increases with decreasing values of the true population ρ values. Another feature is the degrading of the sample $\hat{\rho}$ values that decreases with increasing sample size. In addition, it should be noted that the sample $\hat{\rho}$ values of column C3 must necessarily be equal to 1.0 whenever the sample size equals the number of independent re-

TABLE 5a. Population 5: Contaminated population of 4158 events consisting of the initial population of 3958 events and 200 very extreme events. Columns are (C1) true population ρ values, (C2) average of 10 000 sample $\hat{\rho}$ values based on population regression coefficients, (C3) average of 10 000 sample $\hat{\rho}$ values based on regression coefficients for each sample, and (C4) average of 5 sample $\hat{\rho}$ values for each of 10 000 random sets of regression coefficients associated with the 10 000 samples of C3.

Sample size	Case	Model	C1	C2	C3	C4
15	10	LAD	0.16541	0.15865	0.82410	0.15671
		LSD	0.13645	0.12858	0.75637	0.17684
	6	LAD	0.10284	0.09984	0.65648	0.21212
		LSD	0.08999	0.08536	0.59195	0.23220
25	10	LAD	0.16541	0.16201	0.67046	0.23397
		LSD	0.13645	0.13157	0.61289	0.25467
	6	LAD	0.10284	0.10133	0.52745	0.22739
		LSD	0.08999	0.08691	0.48267	0.24628
40	10	LAD	0.16541	0.16389	0.57754	0.26244
		LSD	0.13645	0.13407	0.53669	0.27945
	6	LAD	0.10284	0.10222	0.42604	0.21735
		LSD	0.08999	0.08863	0.39821	0.23444
65	10	LAD	0.16541	0.16494	0.48744	0.26077
		LSD	0.13645	0.13526	0.46297	0.27581
	6	LAD	0.10284	0.10283	0.33718	0.19921
		LSD	0.08999	0.08927	0.30874	0.20295
100	10	LAD	0.16541	0.16441	0.40913	0.24710
		LSD	0.13645	0.13491	0.38873	0.25578
	6	LAD	0.10284	0.10221	0.27822	0.18593
		LSD	0.08999	0.08906	0.23511	0.17010
160	10	LAD	0.16541	0.16504	0.33696	0.23219
		LSD	0.13645	0.13564	0.30317	0.22288
	6	LAD	0.10284	0.10274	0.23075	0.17241
		LSD	0.08999	0.08959	0.17580	0.14063
250	10	LAD	0.16541	0.16478	0.28582	0.21974
		LSD	0.13645	0.13574	0.23539	0.19305
	6	LAD	0.10284	0.10253	0.19584	0.15837
		LSD	0.08999	0.08960	0.13849	0.11975
500	10	LAD	0.16541	0.16524	0.23315	0.20148
		LSD	0.13645	0.13631	0.17699	0.16198
	6	LAD	0.10284	0.10269	0.15889	0.14052
		LSD	0.08999	0.08981	0.11112	0.10326

gression variables (i.e., $n = p + 1$). As noted in the discussion of population 1, case 10, which involves 10 independent variables, yields a true ρ value that is almost identical to the one for case 6, which uses only 6 independent variables. However, the sample $\hat{\rho}$ values of column C3 in Table 1a are distinctly larger for case 10 than for case 6. Thus, the degrading of the $\hat{\rho}$ values for case 10 is much more severe than for case 6. The rule of parsimony is confirmed here: use the fewest number of independent variables as possible for any given situation. Of course, independent variables that make substantial contributions to the size of $\hat{\rho}$ must be kept; if any of the remaining 6 independent variables of case 6 were to be removed, the true value of ρ would reflect a nontrivial reduction in agreement. With the exception of population 5 in which the performance universally fails, case 6 clearly has an advantage over case 10. Thus, the following discussion related to degrading is restricted to case 6. Sample sizes less than $n = 40$ produce severe degrading of the $\hat{\rho}$ values for both the LAD and LSD analyses. This reduction is likely due to the small ratio of $n - p - 1$ to $p + 1$ (i.e., very little information is available per predictor).

For population 1, which involves no contamination, the degrading associated with the LSD regression model is less than the degrading associated with the LAD regression model in every instance. However, the degrading associated with the LSD regression model is greater than the degrading associated with the LAD regression model for populations 2–5. This feature is further exaggerated for populations 3 and 5, which contain severely contaminated data. A further feature of populations 3–5 is that the average $\hat{\rho}$ values of the LSD analyses exceed the true population ρ values in column C1 (see Tables 3–5). This same disconcerting result is true for both the LAD and LSD regression models in Table 5. Thus it is concluded that both the LAD and LSD regression models fail for those populations containing even 5% severe contamination.

Except for the degrading feature of the LSD analyses documented in Table 2, the LSD regression model appears to do a reasonable job for populations involving small amounts of moderate contamination. If a population contains either small amounts (roughly 1%) of severe contamination or up to approximately 5% of moderate contamination, the LAD regression model is

TABLE 5b. Population 5: Contaminated population of 4158 events consisting of the initial population of 3958 events and 200 very extreme events. Columns are ratio estimators C2/C1, C3/C1, C4/C1, and C4/C3 associated with C1, C2, C3, and C4 of Table 5a.

Sample size	Case	Model	C2/C1	C3/C1	C4/C1	C4/C3
15	10	LAD	0.959	4.982	0.947	0.190
		LSD	0.942	5.543	1.296	0.234
	6	LAD	0.971	6.384	2.063	0.323
		LSD	0.949	6.578	2.580	0.392
25	10	LAD	0.979	4.053	1.414	0.349
		LSD	0.964	4.492	1.866	0.416
	6	LAD	0.985	5.129	2.211	0.431
		LSD	0.966	5.364	2.737	0.467
40	10	LAD	0.991	3.492	1.587	0.454
		LSD	0.983	3.933	2.048	0.521
	6	LAD	0.994	4.143	2.113	0.510
		LSD	0.985	4.425	2.605	0.589
65	10	LAD	0.997	2.947	1.577	0.535
		LSD	0.991	3.393	2.021	0.596
	6	LAD	1.000	3.279	1.937	0.591
		LSD	0.992	3.431	2.255	0.657
100	10	LAD	0.994	2.473	1.494	0.604
		LSD	0.989	2.849	1.875	0.658
	6	LAD	0.994	2.705	1.808	0.668
		LSD	0.990	2.613	1.890	0.723
160	10	LAD	0.998	2.037	1.404	0.689
		LSD	0.994	2.222	1.633	0.735
	6	LAD	0.999	2.244	1.676	0.747
		LSD	0.996	1.954	1.563	0.800
250	10	LAD	0.996	1.728	1.328	0.769
		LSD	0.995	1.725	1.415	0.820
	6	LAD	0.997	1.904	1.540	0.809
		LSD	0.996	1.539	1.331	0.865
500	10	LAD	0.999	1.410	1.218	0.864
		LSD	0.999	1.297	1.187	0.915
	6	LAD	0.999	1.545	1.366	0.884
		LSD	0.998	1.235	1.147	0.929

recommended over the LSD regression model. Because investigators usually know neither the amount nor the severity of contamination for a given population under study, the LAD regression model analysis appears to be the best choice for avoiding potential problems associated with possible population contamination.

The results of this study provide estimates of average degrading from interpolations of the values in Tables 1–5. The fact that the estimates are based on average degrading must be emphasized. For a specified study the actual amount of degrading may vary from none to far more than the average loss of skill since the values vary about the average degrading determined for the population. A final point regarding the results of this study is that the average degrading depends on a specific population. The results are anticipated to be different if other populations are analyzed, even with the same true population ρ values and the same sample sizes used here.

b. Degrading equations

Nonlinear degrading equations are constructed for both the LAD and LSD regression models. These equations yield predicted population ρ values ($\tilde{\rho}$) and are functions of the obtained sample $\hat{\rho}$ values based on the six predictors of case 6 in column C3 of Tables 1–5 and the difference ($w = n - p - 1$) between the sample sizes (n) and the number of unknown parameters ($p + 1$) in the respective regression models. Both equations have the form given by

$$\tilde{\rho} = \min\left\langle \hat{\rho}, \max\left\{0, 1 - \left[\frac{\ln(\hat{\rho})}{\mathrm{H}(w)}\right]^{1/1.32}\right\}\right\rangle$$

and

$$\mathrm{H}(w) = \min(-10^{-50}, \beta_1 w^{0.04} + \beta_2 w^{0.06} + \beta_3 w^{0.08})$$

where, for the LAD regression model,

$$\beta_1 = 147.85585$$

$$\beta_2 = -266.53958$$

$$\beta_3 = 118.99034,$$

and for the LSD regression model,

$$\beta_1 = 155.60230$$

$$\beta_2 = -279.97996$$

$$\beta_3 = 124.79452.$$

TABLE 6. Nondegraded and degraded measures of agreement associated with LAD and LSD regression models. Column headings indicate the seasonal total of NS (named storms), H (hurricanes), HD (hurricane days), IH* (intense or major hurricanes), IHD* (intense hurricane days), HDP* (hurricane destruction potential), and NTC* (net tropical cyclone activity). Note that the data with asterisks are bias corrected following the analysis of Landsea (1993).

	NS	NSD	H	HD	IH*	IHD*	HDP*	NTC*
LAD regression model								
1 December forecast								
nondegraded	0.440	0.514	0.447	0.493	0.473	0.452	0.451	0.547
degraded	0.305	0.407	0.315	0.379	0.352	0.323	0.321	0.450
1 June forecast								
nondegraded	0.514	0.660	0.617	0.703	0.637	0.614	0.709	0.719
degraded	0.381	0.567	0.514	0.617	0.539	0.511	0.624	0.636
1 August forecast								
nondegraded	0.447	0.608	0.472	0.516	0.598	0.579	0.555	0.598
degraded	0.299	0.513	0.335	0.396	0.501	0.478	0.447	0.501
LSD regression model								
1 December forecast								
nondegraded	0.359	0.407	0.388	0.400	0.478	0.428	0.428	0.524
degraded	0.174	0.253	0.223	0.242	0.356	0.284	0.284	0.418
1 June forecast								
nondegraded	0.441	0.625	0.549	0.652	0.590	0.574	0.658	0.672
degraded	0.270	0.520	0.423	0.553	0.476	0.456	0.561	0.578
1 August forecast								
nondegraded	0.435	0.528	0.428	0.489	0.580	0.499	0.523	0.595
degraded	0.277	0.408	0.266	0.355	0.476	0.369	0.402	0.494

As an example for the LAD regression model, suppose $n = 40$, $p = 6$, $w = 33$, and $\hat{\rho} = 0.60$. Then the estimated population ρ value is $\tilde{\rho} = 0.510$. The corresponding estimate for the LSD regression model is $\tilde{\rho} = 0.507$. Note that the LAD ratio $\hat{\rho}/\tilde{\rho} = 0.600/0.510 = 1.177$ and the LSD ratio $\hat{\rho}/\tilde{\rho} = 0.600/0.507 = 1.183$ are estimates of artificial skill and correspond to entries in column C3/C1 of Tables 1–5. Finally, it is emphasized that each estimated population ρ value ($\tilde{\rho}$) is merely a single estimator. If one is fortunate enough to have calculated values close to the population regression coefficients, then obviously, very little degrading will occur. On the other hand, the degrading may be more extreme than indicated by a single estimator. It is not possible to develop similar equations for shrinkage, since C4/C3 is a ratio of two random variables (C3 and C4) and no population values exist for prediction purposes—that is, the values of C1 that were used in the previous prediction models. However, when the population is not contaminated (population 1), it is apparent that the shrinkage is approximately twice the artificial skill. Because artificial skill and shrinkage are of diminished consequence with increasing sample sizes, it is important to update temporal datasets that exhaust all available information. This will serve to increase the sample size and, consequently, decrease both artificial skill and shrinkage.

c. Application to recent studies

The results of this study regarding the degrading of forecast skill, as measured by agreement coefficients, permit clarification and explication of previously reported forecasts. For these purposes, three studies by the authors are used to illustrate the degrading of forecast skill. Specifically, Gray et al. (1992, 1993, 1994) report LAD and LSD regression model nondegraded measures of agreement between various indices of tropical cyclone activity in the Atlantic basin (including the Atlantic Ocean, Caribbean Sea, and Gulf of Mexico): number of named storms (NS), number of named storm days (NSD), number of hurricanes (H), number of hurricane days (HD), number of intense hurricanes (IH), number of intense hurricane days (IHD), hurricane destruction potential (HDP), and net tropical cyclone activity (NTC). The values for IH, IHD, HDP, and NTC have been adjusted to reflect a small overestimation of major hurricane intensities as reported by Landsea (1993) and are identified as IH*, IHD*, HDP*, and NTC* (cf. Gray et al. 1994). These eight indexes of tropical cyclone activity were forecast using both LAD and LSD regression models at three points in time: 1 December, 1 June, and 1 August. The 1 December prediction was based on six predictors (including the intercept) and 41 years of data (Gray et al. 1992); the 1 June prediction was based on 14 predictors and 42 years of data (Gray et al. 1994); and the 1 August prediction was based on 10 predictors and 41 years of data. The 48 nondegraded measures of agreement based on the LAD and LSD prediction models at 1 December, 1 June, and 1 August for the eight indexes of tropical cyclone activity are from Gray et al. (1992, 1993, 1994) and are summarized in Table 6. Also included

in Table 6 are the corresponding 48 degraded measures of agreement.

While it is clearly not possible to apply the degrading formula to a single sample, Table 6 summarizes the C3/C1 degrading (not C4/C3 shrinkage) as applied to the original data of Gray et al. (1992, 1993, 1994) and shows what might vaguely be anticipated, on average, if the Gray et al. (1992, 1993, 1994) studies were repeated many times. Applying the degrading equation to a single sample is analogous to making a statement as to the likely truth of an individual statistical hypothesis. For example, if a type I error is set at $\alpha = 0.05$, a 95% confidence interval is always a statement about the interval not about the population parameter. That is, of all intervals constructed, 95% will contain the true population parameter. However, nothing can be said about a particular interval calculated from a single sample. Any confidence that a researcher has in a particular interval is undefined probabilistically. The same is true of hypothesis testing; no inference can be made about a particular sample. Rejection of a null hypothesis tells us nothing about the particular sample in question. However, in the long run and on average, we can expect to reject the null hypothesis when it is true about 5 in 100 times when $\alpha = 0.05$.

7. The problem of validation

Because artificial skill, such as indexed in the C3/C1 column of Tables 1b, 2b, 3b, 4b, and 5b, is pervasive in most prediction research, investigators often attempt to validate their sample regression coefficients. The usual method is called ''cross-validation'' in which a few observations (up to one-half the observations) are omitted, at random, from a model and the model is then tested on the omitted observations (Michaelsen 1987). More typically, a sample of size n is divided into a construction sample of size $n - 1$ and a validation sample of size 1, and the model is then tested in all possible (n) ways (Stone 1974, 112). In meteorological forecasting research, this cross-validation procedure is usually realized by withholding each year in turn and deriving a forecast model from the remaining $n - 1$ years, checking each of the n forecast models on the year held in reserve (Livezey et al. 1990; Barnston and Van den Dool 1993; Elsner and Schmertmann 1994). Ideally, of course, the construction model based on sample regression coefficients should be validated against several independent validation samples drawn from the population of interest, but few researchers have the luxury of a simulation study in which the entire population is known and available. Column C4 in Tables 1a, 2a, 3a, 4a, and 5a and the C4/C1 column in Tables 1b, 2b, 3b, 4b, and 5b contain the results of just such a validation study in which, it will be recalled, sample regression coefficients are applied to five new independent samples of size n, and $\hat{\rho}$ agreement values between the y and $\tilde{y}$ values are computed for each of the five samples. With 10 000 construction samples and 5 validation samples, each entry in column C4 is the average of 50 000 $\hat{\rho}$ values. It is abundantly clear from even a cursory inspection of the C4/C1 column in Tables 1b, 2b, 3b, 4b, and 5b that when contamination is present the LSD regression model produces inflated estimates of validation fit and expected skill.

Inspection of column C1 in Table 1a reveals that both the LAD and LSD regression models yield about the same amount of agreement between the y and $\tilde{y}$ values in (the noncontaminated) population 1. Although the LAD regression model yields systematically higher agreement coefficients in every instance, the differences are slight. On the other hand, inspection of column C1 in Table 3a reveals that the LAD and LSD regression models yield quite different amounts of agreement between the y and $\tilde{y}$ values in population 3, which contains 1% severe contamination. While the LAD regression model yields slightly lower agreement coefficients when compared to the agreement coefficients in the noncontaminated population 1, the LSD regression model yields greatly reduced agreement coefficients, relative to those obtained in population 1. Because the contaminated data points in population 2–5 were added at the extremes of the independent variables, but cluster around the median of the dependent variable, the added data points exert a considerable amount of both leverage and influence that are magnified by the intrinsic squaring function inherent in the LSD regression model. These additional data points have the general effect of moving the regression plane from its position in the noncontaminated population and increasing the overall sum of squared prediction errors, resulting in a lower coefficient of agreement. The LAD regression model, based on absolute deviations, is less affected by these extreme values, as reflected in its higher agreement values.

When a sample is drawn with replacement from a contaminated population, the chances are that the sample will not contain any of these extreme values, especially when the sampling fraction n/N is small. Thus, considering column C4 of Table 3a where sample-based regression coefficients are validated against five new, independent random samples, it is clear that, in most cases, successive samples are more representative of each other than they are of the population from which they were drawn. This results in higher average sample agreement values than exist in the population for the LSD regression model. The C4/C1 ratios are given in column C4/C1 of Table 3b, where it can be observed that nearly every LSD validation fit exceeds 1.0. The LAD regression model, based on absolute deviations about the median, is relatively unaffected by even 1% severe contamination, but the LSD regression model, based on squared deviations about the mean, systematically overestimates the validation fit and yields greatly inflated indexes (column C4/C1) of expected skill.

Acknowledgments. This analysis was supported by NOAA and NSD Climate Research Grants. We appreciate the encouragement and support of Dr. David Rodenhuis and Dr. Jay Fein.

REFERENCES

Badescu, V., 1993: Use of Willmott's index of agreement to the validation of meteorological models. *Meteor. Mag.,* **122,** 282–286.

Barnston, A. G., and H. M. Van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate,* **6,** 963–977.

Copas, J. B., 1983: Regression, prediction and shrinkage. *J. Roy. Statist. Soc.,* **B45,** 311–354.

Cotton, W. R., G. Thompson, and P. W. Mielke, 1994: Real-time mesoscale prediction on workstations. *Bull. Amer. Meteor. Soc.,* **75,** 349–362.

DeMaria, M., and J. Kaplan, 1994: A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting,* **9,** 209–220.

Elsner, J. B., and C. P. Schmertmann, 1993: Improving extended-range seasonal predictions of intense Atlantic hurricane activity. *Wea. Forecasting,* **8,** 345–351.

——, and ——, 1994: Assessing forecast skill through cross-validation. *Wea. Forecasting,* **9,** 619–624.

Gray, W. M., C. W. Landsea, P. W. Mielke, and K. J. Berry, 1992: Predicting Atlantic seasonal hurricane activity 6–11 months in advance. *Wea. Forecasting,* **7,** 440–455.

——, ——, ——, and ——, 1993: Predicting Atlantic basin seasonal tropical cyclone activity by 1 August. *Wea. Forecasting,* **8,** 74–86.

——, ——, ——, and ——, 1994: Predicting Atlantic basin seasonal tropical cyclone activity by 1 June. *Wea. Forecasting,* **9,** 103–115.

Hess, J. C., and J. B. Elsner, 1994: Extended-range hindcasts of tropical-origin Atlantic hurricane activity. *Geophys. Res. Lett.,* **21,** 365–368.

Holland, G. J., 1993: Tropical cyclone motion. Global Guide to Tropical Cyclone Forecasting, WMO/TC 560, Rep. TCP-31, World Meteorological Organization, Geneva, Switzerland, 363 pp.

Jarvinen, B. R., C. J. Neumann, and M. A. S. Davis, 1984: A tropical cyclone data tape for the North Atlantic basin, 1886–1983: Contents, limitations, and uses. NOAA Tech. Memo. NWS NHC 22, 21 pp.

Landsea, C. W., 1993: A climatology of intense (or major) Atlantic hurricanes. *Mon. Wea. Rev.,* **121,** 1703–1713.

Livezey, R. E., A. G. Barnston, and B. K. Neumeister, 1990: Mixed analog/persistence prediction of seasonal mean temperatures for the USA. *Int. J. Climatol.,* **10,** 329–340.

McCabe, G. J., and D. R. Legates, 1992: General-circulation model simulations of winter and summer sea-level pressures over North America. *Int. J. Climatol.,* **12,** 815–827.

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

Reynolds, R. W., 1988: A real-time global sea surface temperature analysis. *J. Climate,* **1,** 75–86.

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. Roy. Statist. Soc.,* **B36,** 111–147.

Tucker, D. F., P. W. Mielke, and E. R. Reiter, 1989: The verification of numerical models with multivariate randomized block permutation procedures. *Meteor. Atmos. Phys.,* **40,** 181–188.

Watterson, I. G., 1996: Nondimensional measures of climate model performance. *Int. J. Climatol.,* in press.

Willmott, C. J., 1982: Some comments on the evaluation of model performance. *Bull. Amer. Meteor. Soc.,* **63,** 1309–1313.

——, S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'Donnell, and C. M. Rowe, 1985: Statistics for the evaluation and comparison of models. *J. Geophys. Res.,* **90,** 8995–9005.