Linear Prediction of Indian Monsoon Rainfall*

TIMOTHY DELSOLE

Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

J. SHUKLA

George Mason University, Fairfax, Virginia, and Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

(Manuscript received 15 March 2002, in final form 5 July 2002)

ABSTRACT

This paper proposes a strategy for selecting the best linear prediction model for Indian monsoon rainfall. In this strategy, a cross-validation procedure first screens out all models that perform poorly on independent data, then the error variance of every remaining model is compared to that of every other model to test whether the difference in error variances is statistically significant. This strategy is shown to produce better forecasts on average than selecting either the model that utilizes all predictors, the model that explains the most variance in the independent data, or the model with the most favorable statistic suggested by Mallow. All of the model selection criteria suggest that regression models based on two to three predictors produce better forecasts on average than regression models using a larger number of predictors. For the period up to 1967, the forecast strategy selected the prior climatology as the best predictor. For the period 1967 to the present, the strategy performed better than forecasts based on the prior climatology and all other methodologies investigated. Indexes of Pacific Ocean temperature in the Tropics (called Niño-3) and indexes of pressure fluctuations in the Northern Atlantic (called NAO), at 1–6 lead months, failed to provide prediction models that performed better on average than a prediction based on the antecedent climatology. Forecasts based on the prior 25-yr climatology had especially high skill during the recent period 1989–2000, a fact that appears to be a mere coincidence, but which may be relevant to interpreting the skill of the power regression model currently used by the India Meteorological Department.

1. Introduction

The rainfall over India displays a spectacular annual cycle in which more than 80% of the precipitation occurs in the months June–September (JJAS). This cycle, called the Indian summer monsoon, is attributed to the land–sea temperature contrasts that grow as a result of summer heating and represents the most dramatic regional manifestation of the large-scale Asian monsoon system. Although regional rainfall can have large year-to-year fluctuations, the interannual variability of total India rainfall is about 10% of the mean rainfall. These fluctuations can have devastating impacts on the populations of Asia and India by altering the agricultural production and availability of drinking water. Consequently, the problem of predicting the onset, withdrawal, and total amount of monsoon rain is a high priority in many Asian countries. Despite substantial efforts throughout the world, current atmospheric general circulation models cannot realistically simulate, much less predict, the structure and magnitude of the intraseasonal and interannual variability of summer precipitation over the Asian–Australian monsoon region, even with observed sea surface temperatures (Sperber and Palmer 1996). Coupled ocean–atmosphere models show even more deficiencies (Delecluse et al. 1998).

The fact that dynamical models give unsatisfactory monsoon predictions does not preclude the possibility that a statistical model could give useful forecasts. To construct a statistical model, one first postulates an equation relating the forecast variables, then one estimates the parameters in the model in such as way as to minimize the error of the predictions. This assumes that past statistical relations will be maintained in the future.

The statistical prediction of Indian monsoon rainfall has a long and venerable history. Shukla et al. (1986) review much of this history, beginning with the pioneering work of Blanford in the late 1800s and Walker in the early 1900s. Hastenrath (1995) provides a review of recent investigations. Approximately a dozen predictors have emerged from these studies as being im-

* An expanded version of this paper appeared as COLA Technical Report 114.

Corresponding author address: Dr. Timothy DelSole, Center for Ocean–Land–Atmosphere Studies, 4041 Powder Mill Rd., Suite 302, Calverton, MD 20705-3106.
E-mail: delsole@cola.iges.org

portant for predicting India monsoon rainfall. These predictors include 1) location of the 500-hPa subtropical ridge over India, 2) Himalayan snow cover, 3) sea surface temperature, 4) the Southern Oscillation, 5) surface temperature over India, and 6) the surface pressure over the Northern Hemisphere. A very basic question, which does not seem to have been addressed systematically, is which of these predictors should be utilized at a given time?

To address the above question, two problems of statistical prediction need to be recognized. First, even if an equation fits an historical record very well, it may predict new, independent data very poorly. Second, no unique regression equation exists for a fixed set of predictors. Indeed, 2^p distinct regression equations can be derived from a pool of p predictors. Thus, the key problem in statistical prediction is not in constructing regression models, but in *choosing* a good model, out of a vast pool of models, that will give good predictions *in the future*. This is called the problem of model selection. These problems and methods for addressing them are reviewed in section 2.

The primary purpose of this paper is to suggest a solution to the problem of model selection, and to draw conclusions regarding Indian monsoon rainfall prediction based on this solution. We will consider only linear regression models—that is, forecast models that are linearly related to a finite set of predictors, with coefficients obtained by the method of least squares. A strategy for selecting the best model for prediction is proposed in section 3. In this procedure, a cross-validation procedure first screens out all models that are likely to perform poorly on independent datasets, then the prediction error of each model is compared with those of all other models to determine whether the difference in error variance exceeds some threshold of significance. Other strategies, such as choosing the model that explains the most variance in independent datasets, and choosing the model with the most favorable statistic suggested by Mallow, are also considered. The skill of these strategies on monsoon-type datasets are discussed in sections 4 and 5. Implications of these results for forecasts by the power regression model used by the India Meteorological Department are discussed in section 6. The main results of this paper are summarized in the concluding section.

2. Review of linear prediction

In this section we review basic concepts and definitions of linear regression prediction. The first task is to determine a least squares estimate of the JJAS Indian monsoon rainfall, $R(t)$, based on observations of P variables $x_1, x_2, \ldots, x_P$. The variable we want to predict, $R(t)$, is called the predictand, and the antecedent variables on which the prediction is based, $x_1, x_2, \ldots, x_P$, are called the predictors. The prediction equation is assumed to be of the linear form

$$R_p(t) = a_0 + a_1x_1(t - \tau_1) + a_2x_2(t - \tau_2) + \cdots + a_Px_P(t - \tau_P) + \zeta(t), \quad (1)$$

where R_p represents the least squares estimate of R, the parameters $a_0, a_1, \ldots, a_P$ are to be determined from data, the τ values are lead times, and ζ represents the prediction error. Here t is discrete with N distinct values. The random prediction errors $\zeta(t)$ are assumed to be independent, normally distributed with zero mean and constant variance.

The parameters $a_0, a_1, \ldots, a_P$ that minimize the mean-squared errors $(R - R_p)^2$ can be determined by the method of least squares. A useful introduction to this topic can be found in von Storch and Zwiers (1999). The solution is as follows. Suppose (1) is written in the equivalent vector form $\mathbf{R}_p = \mathbf{x}^T\mathbf{a}$, where superscript T denotes a transpose, $\mathbf{x} = (1, x_1, x_2, \ldots, x_P)$, and $\mathbf{a} = (a_0, a_1, \ldots, a_P)$. The constant term has been included in concise form by introducing an additional "predictor" whose value is always unity. In this notation, the least squares solution is given by

$$\mathbf{a} = \langle \mathbf{x}\mathbf{x}^T \rangle^{-1} \langle \mathbf{x}R \rangle, \quad (2)$$

where brackets denote an average over the whole record of N data points. The quality of the fit can be measured by the mean-squared prediction error, e_d, and by the anomaly correlation coefficient, ρ_d:

$$e_d = \langle (R_p - R)^2 \rangle = \langle R^2 \rangle - \mathbf{a}^T \langle \mathbf{x}R \rangle$$

$$\rho_d = \frac{\langle RR_p \rangle}{\sqrt{\langle R^2 \rangle \langle R_p^2 \rangle}} = \sqrt{\frac{\mathbf{a}^T \langle \mathbf{x}R \rangle}{\langle R^2 \rangle}}. \quad (3)$$

An additional metric that will prove convenient is

$$s_d = 1 - \frac{e_d}{\langle R^2 \rangle} = \rho_d^2. \quad (4)$$

This metric indicates the fraction of variance explained by the forecasts. It is bounded above by unity and vanishes when $\mathbf{a} = 0$.

A large value of s_d merely indicates that the equation *fits the data* well, it does not necessarily imply that it *predicts the future* well. To test the predictive performance of the model, we apply a cross-validation procedure, assuming that monsoon rainfall in any given year is statistically independent of the rainfall in other years. To apply this procedure, the data are divided into two disjoint sets: one for estimating (or "training") the model, called the dependent sample, and the other for verifying (or "testing") the forecasts, called the independent sample. Suppose the total record can be partitioned into K mutually exclusive sets of equal size; denote the kth partition by I_k. Any one of these partitions can be considered the independent set, while the complement of this set, D_k, can be considered the dependent set. The a linear prediction model determined strictly from the kth dependent set is $R_p = \mathbf{x}^T\mathbf{b}_k$, where

$$b_k = \left(\sum_{D_k} \mathbf{x}\mathbf{x}^{\mathrm{T}}\right)^{-1}\left(\sum_{D_k} \mathbf{x}R\right). \tag{5}$$

The crux of cross validation is to verify this model strictly against the *independent* set. The quality of the resulting forecasts can be measured by the mean-square (MS) forecast errors and anomaly correlation coefficient (ACC) *in the independent set*:

$$\varepsilon_k = \overline{(R_p - R)^2}^k = \overline{R^2}^k + \mathbf{b}_\mathbf{k}^{\mathrm{T}}(\overline{\mathbf{x}\mathbf{x}^{\mathrm{T}}}^k)\mathbf{b}_\mathbf{k} - 2\mathbf{b}_\mathbf{k}^{\mathrm{T}}(\overline{\mathbf{x}R}^k)$$

$$r_k = \frac{\overline{RR_p}^k}{\sqrt{\overline{R^2}^k\overline{R_p^2}^k}} = \frac{\mathbf{b}_\mathbf{k}^{\mathrm{T}}(\overline{\mathbf{x}R}^k)}{\sqrt{\overline{R^2}^k[\mathbf{b}_\mathbf{k}^{\mathrm{T}}(\overline{\mathbf{x}\mathbf{x}^{\mathrm{T}}}^k)\mathbf{b}_\mathbf{k}]}}, \tag{6}$$

where overbar $\overline{(\cdot)}^k$ denotes an average over the kth *independent* dataset. This procedure can be repeated for each independent dataset. Averaging over all independent sets gives

$$e_i = \frac{1}{K}\sum_{k=1}^{K}\varepsilon_k, \qquad s_i = 1 - \frac{e_i}{\left(\frac{1}{K}\sum_{k=1}^{K}\overline{R^2}^k\right)}, \quad \text{and}$$

$$\rho_i = \frac{\left(\sum_k \overline{RR_p}^k\right)}{\sqrt{\left(\sum_k \overline{R^2}^k\right)\left(\sum_k \overline{R_p^2}^k\right)}}, \tag{7}$$

where e_i is the independent error variance, s_i is the variance explained in the independent set, and ρ_i is the overall anomaly correlation of the predictions in the independent set.

The distinction between dependent and independent error variances, e_d and e_i, becomes clear for small sample sizes N. Lorenz (1977) showed that, if the predictors and predictand are chosen randomly and independently from a normal distribution, in which case the variables have no correlation in the population but could have a large correlation in a small sample, then the expected values of the independent and dependent error variances e_i and e_d are

$$E[e_d] = \left(\frac{N - P - 1}{N}\right)E[R^2]$$

$$E[e_i] = \left(\frac{N - 1}{N - P - 2}\right)E[R^2], \tag{8}$$

where $E[\ ]$ denotes an ensemble average (Lorenz's notation has been adjusted to conform to ours). For large N, the two error estimates approach the correct value of $E[R^2]$. But as P/N increases, the expected error in the dependent data tends to vanish while the expected error in the independent set approaches infinity. Davis (1976) derived similar results in the context of autoregressive models for moderately large N. In essence, if the sample size is small compared to the number of parameters, the parameter values *adapt* to the peculiarities of the sample. As a result, such models perform very well on the dataset from which they were derived, but perform poorly on a new, independent dataset. This problem is called artificial skill. Since sample size is always limited in practice, the existence of artificial skill implies that empirical models should employ the smallest number of predictors for adequate representation. This principle, known as the principle of parsimony, can be justified in other rigorous ways (Box et al. 1994).

The problem with statistical prediction can now be seen clearly: the problem is not in constructing a forecast model, but in choosing a single model that will produce good forecasts on an independent dataset. One cannot simply choose the model that fits the most variance in the available record, because this model would always be the one that includes all possible predictors, even if the predictors have no physical relation with the predictand. Moreover, owing to artificial skill, the number of predictors should be as small as possible. Thus, to decide whether a set of predictors should be used in a regression model, a systematic methodology is needed to balance the negative impact of artificial skill against the positive impact of capturing useful predictive information.

3. Criteria for selecting a regression model

In this section we suggest a method for selecting a prediction model. Consider first the problem of selecting the ''best model'' out of all models with only P predictors. An obvious choice is the model with the least mean-square cross-validated error, e_i, since this model is most likely to perform well on independent datasets drawn from the same probability distribution. To find this model, we first fix the *number* of predictors P, and then search all possible combination of predictors to find the regression model that maximizes s_i. That is, we construct a least squares model for every possible combination of the physical parameters: if $P = 1$, then we consider x_1 alone, x_2 alone, . . . ; if $P = 2$, then we consider x_1 and x_2, x_1 and x_3, . . . ; and so on. This is called the ''all possible regressions procedure.'' Then, the combination of variables yielding the maximum explained variance in the independent sample, s_i, is selected. We call this the *screening procedure*. Note that -for a pool of P predictors there are 2^P regression models—the number of models increases exponentially with P. Thus, screening is not practical if the pool of predictors is large.

The above screening procedure leads to a specific set of predictors for each P. We propose the following strategy for selecting the best value of P. First, we test whether an observed increase in s_d, in going from a P to a $P + M$ parameter model, is larger than would be expected by random chance. If not, then the principle of parsimony calls for rejecting the $P + M$ model in favor of the model with fewer parameters. This test,

which we call the F test, is a standard procedure in linear regression (see von Storch and Zwiers 1999 for an introduction). To understand this test, consider two models of the predictand R:

$$\text{Model A:} \quad R = a_0 + a_1x_1 + \cdots + a_Px_P + \zeta$$

$$\text{Model B:} \quad R = b_0 + b_1x_1 + \cdots + b_Px_P + b_{P+1}x_{P+1} + \cdots + b_{P+M}x_{P+M} + \zeta. \tag{9}$$

Model B contains all of the predictors of model A, plus the additional predictors $x_{P+1}, x_{P+2}, \ldots, x_{P+M}$. We want to know whether the additional predictors in model B contribute substantial predictive information not contained in the variables $x_1, x_2, \ldots, x_P$. Intuitively, if the additional variables do not contribute much predictive information, then the prediction errors of model B should differ little from those of model A. In essence, we want to test the null hypothesis that there is no real difference between models A and B; that is, our null hypothesis is that $b_{P+1} = b_{P+2} = \cdots = b_{P+M} = 0$. If the null hypothesis is true and the errors are normally distributed, then it can be shown that

$$f = \frac{\left(\dfrac{\langle e_A^2\rangle - \langle e_B^2\rangle}{M}\right)}{\left(\dfrac{\langle e_B^2\rangle}{N - M - P - 1}\right)} \tag{10}$$

follows an F distribution with (M, $N - M - P - 1$) degrees of freedom. The quantities $\langle e_A^2\rangle$ and $\langle e_B^2\rangle$ represent the squared error of models A and B averaged over the dependent sample. Large values of F favor rejection of the null hypothesis, indicating that the M additional parameters lead to a significant reduction in the prediction errors, and therefore that model A should be rejected and model B should be adopted.

Note that the above procedure assumes that the P-parameter model is a *subset* of the $P + M$ parameter model. Conceivably, after the screening procedure, the two models could contain completely different predictors. In practice, however, this discrepancy occurs rarely—the P-predictor model usually is a subset of the $P + M$ predictor model. Moreover, results presented in the next two sections suggest that the procedure still gives sensible results even when one model is not a proper subset of the other. Thus, in this paper, we apply the F test to the screened regression models even when the lower-order model is not a proper subset of the full model.

In order to compare F values with different degrees of freedom, it will prove convenient to convert the F values into a significance level α such that

$$\alpha = 1 - \int_0^f F_{M,N-M-P-1}(x)\, dx, \tag{11}$$

where the integrand is the F distribution function with (M, $N - M - P - 1$) degrees of freedom. Small α calls for rejecting the null hypothesis that the error variances are identical and for accepting the alternative hypothesis that the full model is better than the reduced model. The value of α below which we reject the null hypothesis is called the significance level α_c.

The most appropriate value for the significance level α_c is not obvious. It should be recognized that the all possible regressions procedure leads to a bias toward identifying a poor prediction model. To see this, note that if the probability of a type-I error in a single hypothesis test is p, then the probability of an error at least once in m tests is $1 - (1 - p)^m$, which approaches unity in the limit of large m. Thus, a fairly stringent significance level should be chosen, with the level becoming more stringent as the pool of predictors grows. In this paper, we will explore different critical significance levels.

Since more than two models usually will be compared, a well-posed strategy requires specifying both a critical significance level and a contingency table for every possible result. In this paper, we circumvent the need to specify contingency tables by adopting the most conservative rule, namely, by adopting a model with P predictors only if it is significantly better by the F test than *all* other models with fewer predictors.

The above screening procedure and F test criteria are used only for *selecting the predictors*. After the predictors have been selected, the method of least squares is applied to the *full* dataset to obtain the regression coefficients.

It will prove useful to compare the F-test selection criterion to other criteria. A natural idea is to choose the model that explains the most variance in the independent data; that is, the model that maximizes s_i. This criterion, which we call the "maximum s_i criteria," is intuitively appealing because it gives the model that makes the best predictions of independent data. Nevertheless, it is not ideal for two reasons. First, it makes no reference to the actual value of s_i. For example, if the maximum value were negative, indicating that the predictions are systematically worse than forecasts based on climatology, then we probably would conclude that *no* prediction should be attempted. Second, the maximum s_i may not differ in a statistical sense from other observed values of s_i. Unfortunately, the significance test for s_i is not straightforward to calculate from first principles, since this calculation must account for the cross-validation procedure.

Johnson and Wichern (1998) discuss another criterion based on Mallow's C_p:

$$C_p = \frac{\langle e_A^2\rangle}{\left(\dfrac{\langle e_B^2\rangle}{N - M - P - 1}\right)} - [N - 2(P + 1)]. \tag{12}$$

To understand this statistic, note that if the reduced model is adequate, that is, does not suffer from lack of fit,

then $E[\langle e_A^2 \rangle] = (N - P - 1)\sigma^2$, where σ is the standard deviation of ζ in (9). If, in addition, the error variance of the full model is unbiased, then $E[\langle e_B^2 \rangle] = (N - P - M - 1)\sigma^2$. Thus, if the correct model size is near $P + 1$, then C_p will roughly equal $P + 1$, provided the full model is unbiased. Thus, if we plot $(P + 1, C_p)$ for each subset of predictors, then we can identify good models as those with $(P + 1, C_p)$ coordinates near the 45° line. Near the completion of this work, we discovered other criteria, such as Akaike's information criterion, which some authors advocate strongly (Burnham and Anderson 1998). These other criteria will be considered in future work.

4. Forecast models based on Niño-3 and NAO

In this section we apply the procedures outlined in the previous section to construct prediction equations for monsoon rainfall based only on two indexes, Niño-3 and North Atlantic Oscillation (NAO) defined in section 4a. Section 4b demonstrates the problem with selecting the model that maximizes the dependent variance. The results of the screening procedure are discussed in section 4c, and the results of applying the selection criteria are discussed in sections 4d–4e. It is found that the different strategies do not always select the same model. To determine which strategy selects the best forecast model, the strategies are applied to every continuous 25-yr segment of the record, then the selected model is used to predict the immediately following (independent) 26th year. The performance of these models is evaluated in section 4f. The results of using only Niño-3 or NAO as predictors are discussed in section 4g.

a. The data

The predictors used in this section are the monthly mean Niño-3 and NAO indices, for the months December–May prior to the JJAS Indian monsoon. This gives a total of 12 candidate predictors: 2 indexes for each of the 6 lags. By convention, January is said to be at lag −5, February is at lag −4, and so on until May at lag −1. The datasets from which these indices are derived are the following:

1) Total Indian monsoon rainfall during JJAS, estimated from area-weighted observations at 306 land stations uniformly distributed over India, 1871–2000 (Parthasarathy et al. 1995).
2) Monthly mean Niño-3 (Pacific sea surface temperature over 5°S–5°N, 90°–150°W): 1870–1998 (Hadley Centre, United Kingdom); 1950–2000 [the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center, available online at ftp.ncep.noaa.gov].
3) Monthly mean NAO index (sea level pressure difference between Gibraltar and Stykkisholmur, Iceland), 1870–2000 (University of East Anglia, available online at www.cru.uea.ac.uk).

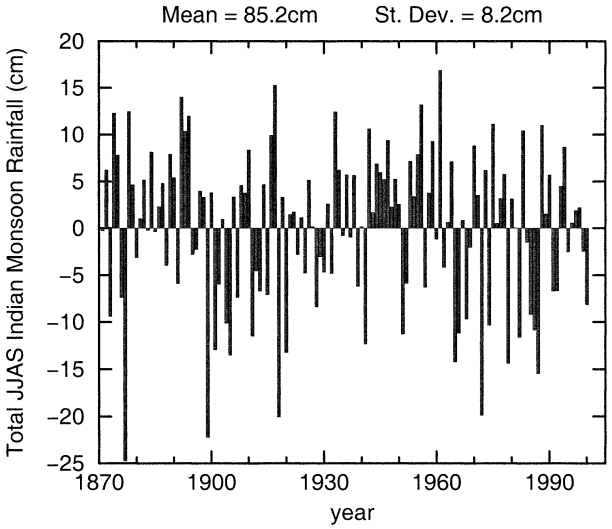

FIG. 1. Total JJAS Indian monsoon rainfall vs yr.

We refer to these data as set I. Results for other predictors are discussed in section 5. For reference purposes, the total Indian monsoon rainfall, minus the grand mean of 85 cm, is shown in Fig. 1. In light of the suggestion by Trenberth (1990) that the climate "shifted" around 1976, the periods 1871–1976 and 1977–2000 will be analyzed separately.

b. Criteria based on maximizing dependent variance

We first show that the model that best *fits* the dependent data is not necessarily the best model for *prediction*. To do this, we fix the number of predictors P, and then search to find the model that best fits the data (i.e., maximizes s_d). Figure 2 shows the variance explained in the dependent sample, s_d, obtained from this search for the period 1977–2000. Note that the variance explained in the dependent sample increases with the number of predictors. On the basis of the solid curve in Fig. 2, one might mistakenly conclude that the "best" prediction model is that which contains all predictors. However, the increase in s_d with the number of predictors is a mathematical certainty, even if the variables are uncorrelated in the infinite ensemble. In contrast, the variance s_i that each model explains in the independent data, as determined by cross validation using continuous 5-yr samples for the independent data, is given by the dashed line in Fig. 2. The figure clearly shows that the variance explained in the independent sample *decreases* as more predictors are added beyond a certain critical value. This result is consistent with the theoretical arguments of Davis (1976) and Lorenz (1977).

c. The screening procedure

Now we apply the screening procedure to find the model that explains the most variance in the independent

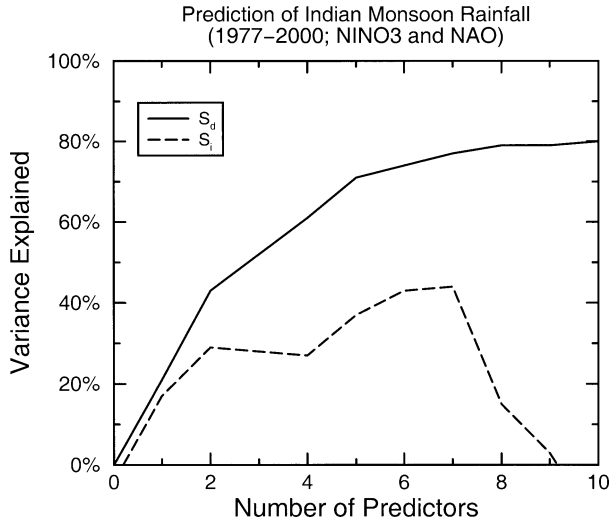

FIG. 2. Variance explained in the dependent (s_d) and independent (s_i) data by the regression model that maximizes s_d. The pool of predictors are the monthly mean Niño-3 and NAO indices at lead months 1, 2, . . . , 6. The regression model is constructed from data in the period 1977–2000.

dataset. To do this, we fix the number of predictors P, then apply the cross-validation method to every combination of predictors to find the model that maximizes s_i at that P. The results for $P = 0$ through 10 for the period 1871–1976 are shown in Fig. 3. The two curves show the results of the cross validation in which the independent datasets consist of 1- and 5-yr periods. The fact that the two curves nearly coincide indicates that the results are not sensitive to the size of the independent dataset. The figure shows a clear peak at two predictors, suggesting that the two-predictor model will produce the best forecasts in independent data. The forecast models are tabulated in Table 1. The table shows that the best two-predictor model consists of Niño-3 at leads −1 and −2, with coefficients of similar magnitude but opposite sign. This result suggests that the best predictor of Indian monsoon rainfall (IMR) during this period is essentially the *tendency* of Niño-3 prior to the monsoon season, plus some additional weighting on the Niño-3 value at lead −1. A similar conclusion was reached by Shukla and Paolino (1983) and Shukla and Mooley (1987) on the basis of a compositing technique. The present result not only confirms this earlier result, but adds to it by demonstrating that there is no better predictor within this dataset.

TABLE 1. Regression models that maximize the variance explained in the independent data in the period 1871–1976, for one, two, and three predictors. Also tabulated are the "goodness of fit" statistics s_d and s_i, which measure the variance explained in the dependent and independent datasets, respectively. The regression coefficients appear to the left of the predictor, and the time lag (in months) appears to the right in parentheses.

Forecast model for JJAS Indian monsoon rainfall (cm)	s_d	s_i
−4.5Niño-3(−1)	10%	7%
+7.3Niño-3(−2) −10.3Niño-3(−1)	17%	12%
−0.3NAO(−5) +7.2Niño-3(−2) −10.3Niño-3(−1)	17%	12%

The same analysis was repeated for the much shorter period 1977–2000. The results are shown in Fig. 4. The similarity of the two curves for the 1- and 5-yr inde-

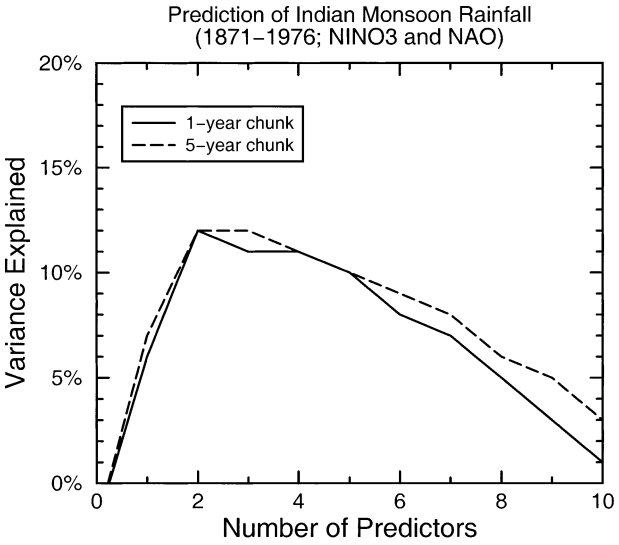

FIG. 3. Variance explained in the independent data by the regression model that maximizes s_i for fixed number of predictors. The pool of predictors are the monthly mean Niño-3 and NAO indexes at lead months 1, 2, . . . , 6. The regression model is constructed from data in the period 1871–1976. The two curves give the results using 1- (solid) and 5-yr (dash) "chunks" for the size of the independent dataset.

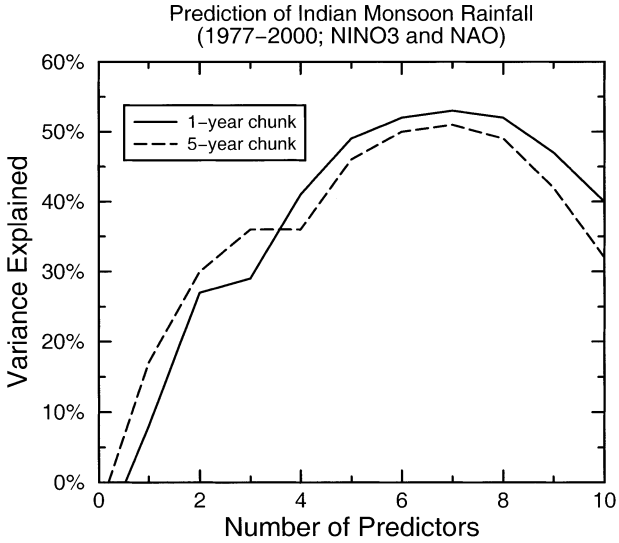

FIG. 4. Variance explained in the independent data by the regression model that maximizes s_i. The pool of predictors are the monthly mean Niño-3 and NAO indices at lead months 1, 2, . . . , 6. The regression model is constructed from data in the period 1977–2000. The two curves give the results using 1- (solid) and 5-yr (dash) chunks for the size of the independent dataset.

TABLE 2. Regression models that maximize the variance explained in the independent data in the period 1977–2000, for one, two, three, four, and five predictors. The format of the table is the same as in Table 1.

Forecast model for JJAS Indian monsoon rainfall (cm)	s_d	s_i
−2.7 −2.7NAO(−1)	21%	17%
−4.8 +1.9NAO(−3) −3.3NAO(−1)	37%	30%
−3.5 −0.9NAO(−6) +1.9NAO(−5) −2.9NAO(−1)	49%	36%
4.2 −0.9NAO(−6) +1.8NAO(−5) −1.4NAO(−2) −3.4NAO(−1)	57%	36%
−2.9 −1.0NAO(−6) +2.1NAO(−5) −2.9NAO(−1) +3.0Niño-3(−6) −5.3Niño-3(−2)	69%	46%

pendent sets again confirms that the results are insensitive to the size of the independent data. We see, however, two major differences from Fig. 3: the explained variances are much higher, and the maximum explained variance occurs at $P = 7$. While the enhanced fractional variance is a plausible consequence of analyzing a shorter absolute period, its magnitude appears too large to be dismissed as sampling errors. For instance, analysis of other 25-yr subsets, such as 1901–25, 1926–50, 1951–75, give variances well below the variances shown in Fig. 4. The corresponding regression models for the period 1977–2000 are tabulated in Table 2 for one to five predictors. A major difference with the analogous result for the period 1871–1976 (Table 2) is that the NAO index arises as a significant predictor. The sign and magnitude of the coefficients do not suggest any simple interpretation in terms of derivatives or running averages of the NAO.

d. The F-test criteria

We now discuss results of the F-test criteria. In this section, we fix the critical level α_c at 1.5% and use only continuous 5-yr samples for the independent data.

The results for the period 1871–1976 are given in Table 3. This table gives the significance level α of the difference in error variances between the screened regression models. The table is structured so that the number of predictors in the reduced model increases to the right, and the number of predictors in the full model, to which the reduced model is compared, increases downward. First note that the first five rows in the zero-predictor column of the reduced model are all less than 0.1%. These small values indicate that screened models with one to five predictors have significantly less error variance than a forecast based on climatology. Thus, we accept the alternative hypothesis that one or more of these predictors significantly reduces the forecast error of the regression model and consider the one-predictor model. The first three significance levels in the one-predictor column of the reduced model are 0.4%, 1.3%, 2.1%, and increases rapidly after that as the number of predictors in the full model increases. Since the suggested critical level is 1.5% and the first two entries are less than this, we conclude that the two- and three-predictor models have significantly different error variances from the one-predictor model. Thus, we continue onward and consider a two-predictor model. The smallest significance level in the two-predictor column is 44.9%, which is large compared to 1.5%. Furthermore, all the significance levels below and to the right of the two-predictor column are large compared to 1.5%. We conclude that the three- or higher-predictor models do not have significantly different error variances than the two-predictor model. Therefore, according to the f-test criterion with $\alpha_c < 1.5\%$, we should select the two-predictor model from this dataset, consistent with the conclusion reached on the basis of Fig. 3.

We have performed a similar analysis for the period 1977–2000. We find that the F test selects the zero-predictor model for a critical level of 1.5%, but selects the five-predictor model for a critical level in the range 2%–3%. These conclusions will be examined more closely in section 4f.

e. Other selection criteria

We now consider alternative selection criteria. If the criteria is to select the model that maximizes the independent variance s_i, Fig. 3 would imply that the two-predictor model derived from the 1871–1976 record

TABLE 3. Significance level of the difference in error variance between the reduced regression model and the full regression model for the period 1871–1976. A small value (say less than 1.5%) calls for rejecting the reduced model in favor of the full model.

Predictors in full model	No. of predictors in the reduced model						
	0	1	2	3	4	5	6
1	0.001	NA	NA	NA	NA	NA	NA
2	0.000	0.004	NA	NA	NA	NA	NA
3	0.000	0.013	0.499	NA	NA	NA	NA
4	0.000	0.021	0.459	0.295	NA	NA	NA
5	0.001	0.029	0.449	0.335	0.297	NA	NA
6	0.002	0.056	0.621	0.537	0.582	1.000	NA
7	0.004	0.094	0.741	0.686	0.756	0.947	0.742
8	0.007	0.146	0.834	0.800	0.868	0.979	0.910
9	0.013	0.213	0.903	0.886	0.938	0.995	0.978
10	0.018	0.258	0.914	0.899	0.940	0.983	0.951

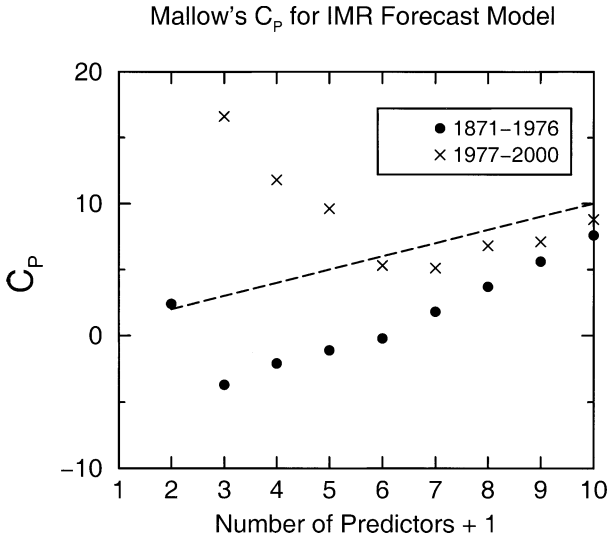

FIG. 5. Mallow's C_p statistic for the screened regression models of Indian monsoon rainfall for the period 1871–1976 (dots) and 1977–2000 (x), and the 45° line (dash).

should be selected, and Fig. 4 would imply that the seven-predictor model derived from the 1977–2000 record should be selected.

The values of Mallow's C_p statistic for the screened regression models are shown in Fig. 5. Recall that the goal is to select the model that lies the farthest below the 45° line, shown as the dashed line. For the 1871–1976 data, the point that lies the farthest below the 45° line is the one for which the number of predictors plus one is three, and therefore the two-predictor model has the most favorable C_p statistic. This selection is consistent with previous conclusions. (The abscissa shows one plus the number of predictors, as is conventional for this procedure.) For the 1977–2000 data, the conclusion is that the five- or six-predictor model has the most favorable C_p statistic. The principle of parsimony implies that the model with fewer predictors, namely five, should be selected.

f. Comparison of selection criteria

At this point we have examined several different selection criteria and obtained the following results. For the period 1871–1976, every criteria selected the two-predictor model. For the period 1977–2000, the *F*-test criteria selected the five-predictor model for $\alpha_c = 3\%$, but selected the zero-predictor model for $\alpha_c = 1.5\%$. On the other hand, Mallow's C_p statistic selects the five-predictor model, while the model that explains the most variance in the independent data is the seven-predictor model. Which of these models should be chosen?

To gain insight into this question, we conducted the following experiment. For each 25-yr segment in the 130-yr record, all possible regression models were calculated, screened, and selected according to a specific criterion. Then, based on the selected P predictors, a regression model was derived from the 25-yr record and used to predict the immediately following (independent) 26th year. For example, a forecast for 1901 was generated by applying this procedure strictly to the 1876–1900 record; and so on. Repeating this procedure for all years 1901–98 gives a time series of independent forecasts that can be used to assess the different selection criteria.

TABLE 4. The rmse and ACC of regression models selected by different criteria for dataset I and the period 1901–98. The forecast in each year was produced by a regression model derived strictly from the prior 25-yr data.

Selection criteria	rmse (cm)	ACC
F test ($\alpha_c = 1.5\%$)	8.8	−0.05
Mallow's C_p	8.9	0.17
Max s_i	10.4	0.06
All predictors	12	0.15
Climatology	8.2	−0.02

The root-mean-square error and anomaly correlation of the forecasts for other selection criteria are given in Table 4. The table shows that while the *F*-test criterion yields the least mean-square error of all selection criteria, its error variance is larger than that of a forecast based on a 25-yr running climatology. Since the regression model depends on the critical significance level chosen for the *F*-test criterion, the possibility exists that a more stringent significance level may lead to better regression models. However, at the 1.5% level used here, only 22 out of the 98 forecasts differ from climatology. Choosing an even more stringent significance level would result in even fewer forecasts differing from climatology. Indeed, for this pool of predictors, it can be verified that *no* critical significance level will produce forecasts better than climatology. Other criteria such as Mallow's C_p statistic, or the maximum s_i, produce forecasts with even larger error variances than the forecasts based on the *F*-test criterion. The fact that none of the selection criteria yield forecasts better than a running climatology suggests that antecedent, monthly mean NAO and Niño-3 indices, *by themselves,* provide little or no predictive information of IMR beyond the climatological mean, in either a mean-square sense or an anomaly correlation sense.

g. Forecasts based on a single physical predictor

The above conclusion seems to contradict the prevailing opinion that Niño-3 preceding the monsoon season is an important predictor of IMR. It is, therefore, of interest to examine the above results more fully. Before doing this, it is worth noting that the prevailing view is based partly on the existence of a significant, *simultaneous* correlation between monsoon rainfall and

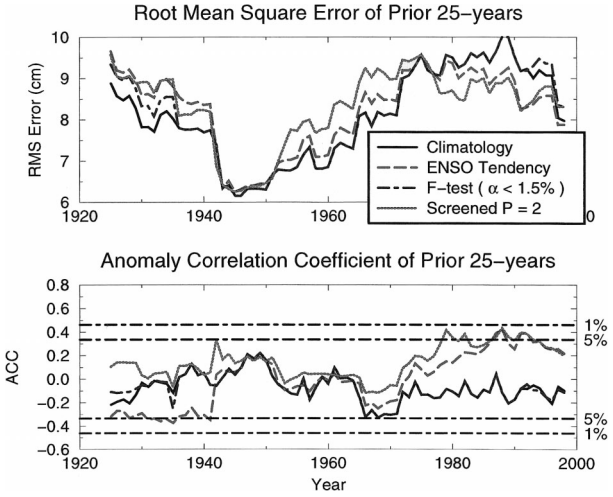

FIG. 6. (top) The 25-yr running rmse and (bottom) 25-yr running correlation between observation and forecast, for four different forecasts: forecast based on prior 25-yr mean (solid), regression forecast based only on the ENSO tendency index (long dash), forecast selected by the F test with $\alpha_c = 1.5\%$ (dash–dot), and forecast by the two-predictor model that maximizes the independent variance (dotted). The forecasts are produced by regression models derived strictly from the 25-yr record prior to the year of forecast. The 1% and 5% significance level for a correlation based on 25 degrees of freedom are shown as straight dash–dot lines in the bottom figure. The F test and climatology curves coincide for the period 1936–92.

ENSO indices (Webster et al. 1998; Kirtman and Shukla 2000). Indeed, the simultaneous 25-yr correlation between JJAS IMR and JJAS Niño-3 has been negative and statistically significant at the 1% level throughout the entire record, except during the 1920s and 1930s, and except for the last 25-yr period ending in 1998. However, the mere fact that a *simultaneous* correlation is significant does not at all imply that the *time-lagged* correlation is significant.

To clarify these issues, we ignore the NAO index for the moment and construct regression models of monsoon rainfall based only on ENSO indices. Furthermore, to connect more closely with previous studies, we consider two new regression models: a regression model derived from a single predictor called ''ENSO tendency,'' which is the March–May (MAM) average minus the December–February (DJF) average Niño-3.4 index (used in section 5 as dataset c), and the regression model based on the two lags of the Niño-3 index that maximize s_I. As before, we construct a regression model from every continuous 25-yr segment in the record, and then use the model to predict the immediately following 26th year. Moreover, instead of presenting the mean-square error and anomaly correlation for the full period 1901–98, we calculate these quantities using a sliding 25-yr window.

Figure 6 shows a 25-yr running rms error and anomaly correlation coefficient (ACC) of the four forecast models. We see that all of the models have a running rmse within 10% of each other. However, the rmse of the climatology forecast is consistently *less* than that of the other forecasts in the period 1925–78, implying that none of the forecasts can beat climatology during this period. For the later period 1979–98, we see that two of the models, namely, the screened two-predictor model and the ENSO-tendency model, have rmse that are on average less than those based on climatology. The corresponding anomaly correlation of the forecasts over a sliding 25-yr period shows that the correlations are positive and significant at the 5% level only in last decades, and that none of the anomaly coefficients are significant at the 1% level. These results support the conclusion that, except in the last two decades, Niño-3 provides little predictive information beyond the climatological mean. The conclusion in the previous section that Niño-3 indices one to six months prior to the monsoon are *not* useful predictors, based on the skill over the full period 1901–98, still is valid *for this period,* since the detectable skill in the last two decades is not sufficiently strong and/or long to dominate the average.

As noted earlier, no *single* value of the critical level produces forecasts better than climatology over the *entire* period. We have verified, however, that a larger value of the critical level (2%–4%) produces forecasts better than climatology *over the last two decades*.

Now we consider the predictive value of the NAO indices of the prior 6 months. For this purpose, two sets of predictors are considered: the predictors selected by the F test with $\alpha_c = 3\%$, and the single predictor that maximizes s_I (the independent variance) in the prior 25-yr period. The forecast skill of the respective models, measured by the 25-yr running rms forecast error and ACC, is shown in Fig. 7. The figure shows that none of the forecasts perform consistently better than a forecast based on the climatology. We have verified that no value of α_c in the F test changes this conclusion. These results lead to the conclusion that, for JJAS monsoon rainfall, the NAO indices of the prior 6 months have little or no predictive value beyond the climatological mean.

5. Forecast models based on upper-level data, land surface data, and other predictors

As mentioned in the introduction, several climate variables aside from the NAO and Niño-3 indices have been suggested as useful predictors of Indian monsoon rainfall. Therefore, it is of interest to repeat the analysis of the preceding section using an expanded pool of predictors. This section presents the results of this analysis using the following predictors, which we call predictor set II, which are all prior to JJAS monsoon:

1) Darwin sea level pressure tendency (Dtend): MAM average minus DJF average, 1883–2000.
2) Darwin sea level pressure tendency: April average minus January average, 1883–2000.

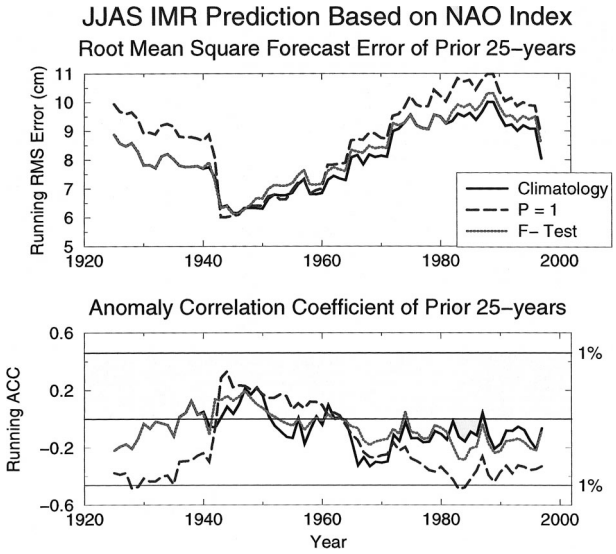

FIG. 7. (top) The 25-yr running rmse and (bottom) correlation between observation and forecast, for three different forecasts: forecast based on prior 25-yr mean (solid), forecast by the one-predictor model that maximizes the independent variance (dash), and forecast selected by the f test with $\alpha_c = 3\%$ (dotted). The forecasts are produced by regression models derived strictly from the 25-yr NAO record prior to the year of forecast. The 1% significance level for a correlation based on 25 degrees of freedom is also shown.

3) Niño-3.4 (5°S–5°N, 170°–120°W) tendency: MAM average minus DJF average, 1870–2000 (Ntend).
4) Niño-3.4 tendency: April average minus January average, 1870–2000.
5) NAO (sea level pressure difference between Gibraltar and Stykkisholmur, Iceland): January–February mean, 1870–2000 (NAO-JF).
6) NAO: April–May mean, 1870–2000 (NAO-AM).
7) Indian surface temperature (25°–35°N, 55°–75°E): MAM average, 1950–95 (TI).
8) Eurasian surface temperature (60°–70°N, 30°–50°E): DJF average, 1950–95 (TE).
9) 500-mb ridge position at 75°E: April average, 1950–2000 (Ridge).

It should be noted that the first six predictors listed above still represent ENSO and NAO indices, but the last three predictors are somewhat independent. We compare the model selection criteria as in section 4f, using $\alpha_c = 3\%$ for set II.

Table 5 shows the rmse and anomaly correlation of the forecasts for the period 1901–98 based on different selection criteria. The table shows that the F-test criteria produces the best forecast in both rmse sense and ACC sense. It also reveals that other criteria produce forecasts that are worse on average than a prediction based on the prior 25-yr climatology. Since only the F-test criteria performs better than climatology, this example clearly demonstrates the need for a good selection criteria.

The 25-yr running rmse and ACC for the F-test models are shown in Fig. 8. Interestingly, the F test selected

TABLE 5. The rmse and ACC of regression models selected by different criteria for dataset II and the period 1901–98. The forecast in each year was produced by a regression model derived strictly from the prior 25-yr data.

Selection criteria	rmse (cm)	ACC
F test ($\alpha_c = 3\%$)	7.6	0.33
Mallow's C_p	8.3	0.22
Max s_i	8.7	0.16
All predictors	8.6	0.21
Climatology	8.2	−0.02

the zero-predictor model for each year in the period 1901–67, implying that no model in this period performed significantly better than a forecast based on the climatology of the prior 25 yr. This idea is supported by the fact that the other selection criteria always selected $p \geq 1$ in this period, but performed worse than climatology. The rmse and ACC of the forecasts for the 1971–98 period, in which the F test always selected $p \geq 1$, are shown in Table 6. The table reveals that all selection criteria performed better than climatology, in both an rmse sense and an ACC sense, but that the F test still performed the best out of all selection criteria. These results suggest that the F-test criteria provides a promising basis for linear prediction.

The above results differ dramatically from those of the previous section. The previous section showed that for dataset I *no* selection criteria gave better forecasts than a running climatology, whereas the present section finds that for dataset II *all* selection criteria beat a running climatology for a limited period. To understand

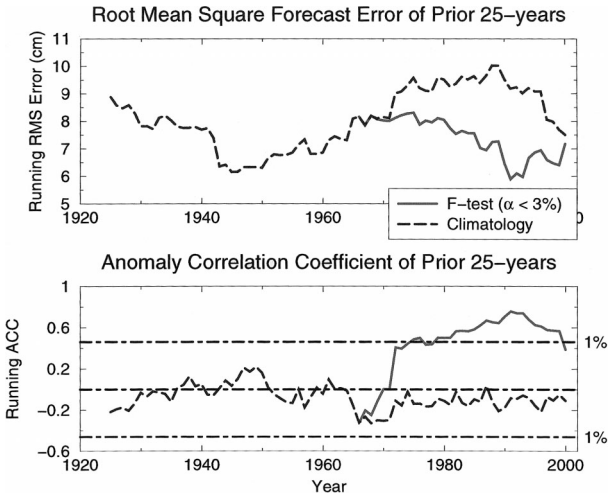

FIG. 8. (top) The 25-yr running rmse and (bottom) correlation between observation and forecast, for regression model selected by the F test ($\alpha_c < 3\%$) (solid) and for forecast based on the 25-yr prior mean (dash). The regression model was constructed from the 25-yr record prior to the year of forecast. The 1% and 5% significance level for a correlation based on 25 degrees of freedom are shown as straight dash–dot lines in the bottom figure. The F test and climatology curves coincide for the period 1925–67.

TABLE 6. The rmse and ACC of regression models selected by different criteria for dataset II and the period 1971–98. The forecast in each year was produced by a regression model derived strictly from the prior 25-yr data.

Selection criteria	rmse (cm)	ACC
F test ($\alpha_c = 3\%$)	6.5	0.62
Mallow's C_p	8.4	0.42
Max s_i	7.4	0.52
All predictors	7.9	0.48
Climatology	8.6	−0.17

this difference, consider the regression equations selected by the F-test criteria for the period 1967–2000 shown in Table 7. The regression coefficients are tabulated under the ''Predictors'' column, with blank entries indicating that the particular predictor was not selected by the F test. The table shows that the preferred predictors are not ENSO and NAO indices *alone*. Rather, the F-test criteria favored three predictors during the period 1967–92, namely, the ridge location, Darwin tendency, and European surface temperature. The years 1993–96 appear to be a transition period with relatively low skill. After 1996 the F-test criteria favored the NAO index as the sole predictor. That the shift occurs in the mid-1990s and the regression model is based on the prior 25 yr suggests that the statistical relations between monsoon rainfall and climate indices shifted in the early 1970s.

Is the skill of the models given in Table 7 dominated by one or two predictors? To answer this question, consider the period 1975–92, in which the available pool of predictors is steady and the models selected by the F test are similar. Moreover, consider just the ridge location, Darwin tendency, and European surface temperature, which are the only indices selected by the F test during this period. We found that, in the case of one predictor, forecasts based on the ridge location performed the worst out of the three, and those based on European surface temperature performed the best. In fact, forecasts based on European surface temperature, alone, performed about as well as the best forecasts based on the optimal combination of the ridge location and Darwin tendency. This greater predictive usefulness

TABLE 7. The forecast model, the forecast, its error, and goodness of fit statistics for the regression model in each year. The predictors were selected by the F test ($\alpha_c < 3\%$) and the regression coefficients were derived from the 25-yr record prior to the year of forecast. The error column was computed as prediction minus observation.

Year	Error (cm)	Forecast (cm)	s_i (%)	s_d (%)	Predictors						
					Constant	Ridge	Dtend	TE	Ntend	NAO-AM	NAO-JF
1967	−2.5	83.6	2	20	98.0		−4.1				
1968	11.9	87.4	−15	0	87.4						
1969	−0.2	82.9	1	21	97.7		−4.1				
1970	−4.0	90.0	0	22	97.4		−4.1				
1971	−3.5	85.2	15	24	97.9		−4.3				
1972	10.9	76.2	17	28	99.0		−4.8				
1973	6.4	97.8	30	41	100.8		−5.7				
1974	6.4	81.2	34	40	99.3		−5.3				
1975	−3.5	92.8	55	62	55.9	2.4	−3.2				
1976	−0.4	85.3	60	64	55.1	2.5	−3.2				
1977	−10.2	78.2	56	62	53.4	2.6	−3.4				
1978	−4.1	86.8	50	57	57.5	2.4	−3.3				
1979	7.8	78.6	51	56	61.4	2.2	−3.6				
1980	−3.4	84.9	52	58	56.5	2.5	−3.5				
1981	−0.3	84.9	55	59	56.3	2.5	−3.5				
1982	1.1	74.7	52	57	58.7	2.3	−3.5				
1983	−6.6	88.9	55	60	56.6	2.4	−3.3				
1984	4.1	87.7	64	72	64.2	2.0	−3.0	1.3			
1985	3.9	79.8	59	70	63.4	2.0	−3.0	1.2			
1986	3.3	77.6	60	71	62.2	2.1	−3.0	1.3			
1987	4.6	74.4	65	77	61.2	2.1	−3.0	1.0			
1988	−10.4	85.7	67	79	61.7	2.1	−3.0	1.1			
1989	5.4	92.0	69	76	63.3	2.1	−4.0	1.2			
1990	−5.1	85.8	69	78	48.9	2.4	−6.0	1.3			
1991	3.7	82.2	68	77	62.4	2.1	−4.0	1.1			
1992	8.3	86.8	65	76	62.6	2.1	−4.0	1.0			
1993	−10.1	79.6	59	77	64.5	2.1	−4.0	1.7			−1.9
1994	−14.7	79.1	62	72	49.0	2.4		1.3	−5.0		
1995	9.3	91.9	51	64	55.9	1.9		1.3	−6.0		
1996	−6.1	79.6	38	48	44.8	2.7			−5.5		
1997	1.0	88.1	26	35	81.1					−6.0	
1998	−1.2	86.2	11	29	81.9					−5.2	
1999	−4.5	78.2	6	28	81.9					−5.1	
2000	16.6	93.7	33	51	81.0					−5.4	1.9

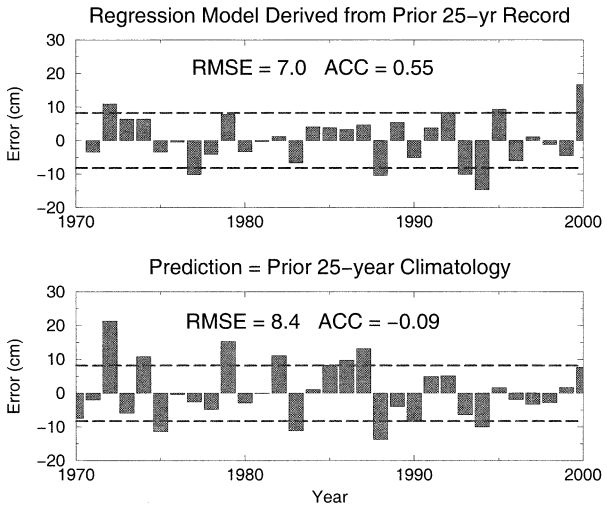

FIG. 9. (top) Forecast error of the regression model selected by the F test ($\alpha_c < 3\%$), and (bottom) forecast error of a prediction based on the mean of the prior 25 yr. The regression model was constructed from the 25-yr record prior to the year of forecast. The dashed line shows the std dev of the monsoon rainfall.

of surface temperature is consistent with the physics of monsoons being driven by the land–sea temperature contrasts. Forecasts with all three predictors performed much better than forecasts based on any two of the three predictors. Thus, the predictive skill of the three-predictor model is not dominated by a single predictor. This result suggests that the climate indices are not necessarily useful predictors *by themselves,* but rather that the indices are useful *in combination with other predictors*. This result is not unexpected, since the F test selects P predictors only if these predictors explain significantly more variance than any model with fewer predictors.

Table 7 and Fig. 9 reveal that prediction models can give errors that exceed the standard deviation of monsoon rainfall. Such large errors occurred 8 out of 26 yr, which is the expected frequency for normally distributed errors with standard deviation equal to the rmse of 7 cm. Other selection criteria produce even higher frequencies. Unfortunately, Table 7 suggests little basis for anticipating the timing of large errors. For instance, large errors occur for both large and small values of s_d and s_i.

It is probably worth pointing out that, for the values of the critical significance level α_c found to be useful, the F-test criteria rarely selects more than three predictors.

6. The 1989–2000 forecasts by the India Meteorological Department

The prediction models explored in this paper differ dramatically from the models currently used by the India Meteorological Department (IMD). Since 1989, the IMD has issued forecasts of monsoon rainfall based on the power regression model proposed by Gowariker et al. (1989, 1991) and Thapliyal and Kulshrestha (1992). This model has the form

$$\frac{R + \alpha_0}{\beta_0} = C_0 + \sum_{i=1}^{16} C_i \left(\frac{X_i + \alpha_i}{\beta_i} \right)^{p_i}, \qquad (13)$$

where R represents the Indian monsoon rainfall, X_i represents the ith physical parameter, and the αs, βs, ps, and Cs are constants chosen to produce good forecasts on historical data. In contrast to the models examined in this paper, the power regression model is nonlinear and utilizes a relatively large number of physical predictors (namely, 16). Our results, however, suggest that the use of such a large number of predictors ought to lead to artificial skill and poor forecasts of independent data. Given the social and economic importance of monsoon forecasts, it is of interest to examine this model more closely.

First note that model (13) contains 49 independent parameters. To see this, note that the βs in (13) can be absorbed with the Cs without loss of generality, leaving a total of three parameters per predictor, plus one constant. Gowariker et al. (1989) estimated the value of the 49 parameters from 37 yr of data. In principle, this model could fit 37 yr of data perfectly. These considerations leave no doubt that the power regression model as used by the IMD is subject to artificial skill.

Given the above considerations, how is it that the final forecast by the IMD has proven ''reasonably accurate''? We cannot address this issue comprehensively because we do not have access to the actual data used by the IMD to produce their forecasts. It should also be noted that the total Indian rainfall dataset used in this paper, based on the data of Parthasarathy et al. (1995), differs slightly from that of the IMD in that the latter includes the hilly and island areas. Nevertheless, we believe that one clue to the answer is the following. Figure 10 shows a 10-yr running rmse of two forecasts: 1) a forecast by the F-test model derived strictly from the prior 25-yr record of dataset II, and 2) a forecast equal to the mean of the prior 25-yr record. Importantly, the climatology forecast performed *unusually* well in the last 5 yr—the MS error of the climatological forecast for the period 1989–2000 is 5.5 cm—whereas for any 10-yr period in 1970–95 the error is 7.5 cm or more (see Figs. 9 and 10), reflecting the fact that monsoon rainfall has been ''normal'' (within 10% of the mean) since 1989.

The above results may help explain the reasonably accurate forecasts by the IMD. As stated in Thapliyal and Kulshrestha (1992), the final forecast by the IMD is not based solely on the power regression model, but rather on a weighted combination of different forecasts. The precise details of this procedure do not appear to be available in the published literature. Nevertheless, if the procedure effectively adjusts the forecast toward the climatology of the immediately preceding period, then

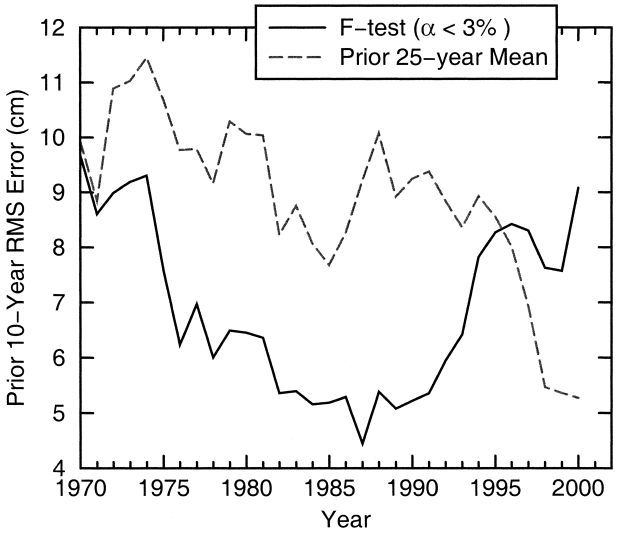

FIG. 10. The 10-yr running rmse for the regression models selected by the *F*-test criteria (solid), and of the errors of the forecast based on the prior 25-yr climatology (dash).

this adjustment will improve the forecast in a mean-square sense in the special period 1989–2000, because in this period a running climatology happens to give a good forecast. That this adjustment occurs is consistent with the fact that the IMD predicted normal rainfall (i.e., within 10% of the mean) for all years during 1989–2000, even though the model has a high probability of extreme forecasts due to nonlinearity and the large number of predictors.

7. Summary and conclusions

The fundamental basis of statistical prediction is that time-lagged correlations observed in the past will persist into the future. In practice, however, the unknown ensemble averages required in statistical regression must be estimated from a finite sample. As a result, statistically derived prediction equations are subject to sampling errors that increase with the number of predictors on which the regression equation is developed. These sampling errors lead to the impression that the regression model fits the data better and better as more predictors are used. However, the forecaster's goal is not to fit the dependent data, but to predict new, independent data. Even if the predictand/predictor correlation estimated from a finite sample is relatively high, the predictive skill of models with many predictors can be very low, perhaps even worse than a prediction based on the prior climatology.

This paper proposed a strategy for selecting the best linear prediction model. In this strategy, all possible prediction models of fixed order are calculated and cross validated to determine the model that minimizes the mean-square error in the independent dataset. This procedure, called the screening procedure, essentially eliminates the models at each order that are not likely to perform well on a new, independent dataset. Then, the mean-square prediction errors of every model are compared with those of every other model to test whether the difference in error variances is statistically significant. This test is applied to every pair of screened regression models, starting with a zero-predictor model (i.e., a forecast based on climatology) and progressing toward increasing number of predictors, until no significantly better model can be found.

To test whether the above strategy can produce useful forecast models, the procedure was applied to every continuous 25-yr segment in the monsoon rainfall record, then the resulting model was used to predict the immediately following (independent) 26th year. The strategy produced better average forecasts for all periods and for all predictor subsets than all other nontrivial methods investigated (e.g., choosing either the model with the most favorable C_p statistic, the model that explains the most variance in the independent data, or the model that utilizes all possible predictors). If the predictors are restricted to ENSO and NAO indices, then none of the strategies selected models that perform consistently better than a prediction based on the climatology of the prior 25 yr, except in the period 1975–2000. We find no evidence to suggest that the ENSO and NAO indices, *by themselves,* are useful predictors of Indian monsoon rainfall prior to 1950 (25 yr prior to 1975). This conclusion covers most finite difference or smoothed versions of these indices, since the pool of predictors included these indices at six different lead times. If, however, the pool of predictors are augmented to include upper-level (500-hPa ridge location) and land surface data (DJF Eurasian temperature), then all of the strategies select models that perform better than climatology, with the screening and the *F*-test criteria performing best of all. Importantly, every forecast model investigated had at least a 20% probability of large error (an error exceeding the standard deviation of the monsoon rainfall).

Our results give little support to the idea that a large number of predictors should be used for long periods of time. None of the model selection criteria investigated here indicate the use of more than two to three predictors, and all of these criteria produced better forecasts on average than the regression models that utilized all the predictors. Also, the screening and *F*-test procedure frequently selects climatology (zero predictors) over any forecast derived from the predictors. Finally, the *F*-test criterion usually selects fewer parameters than a criterion based on maximizing the independent variance, yet the *F* test performs better on average (cf. "*F* test" with "maximum s_i" in Tables 4–6).

On the basis of the available record published in scientific journals, we argued that the power regression model used by the IMD, which is nonlinear and utilizes 16 physical predictors and 49 independent parameters, is subject to artificial skill and has not been proven to

be clearly superior to linear models with a few predictors. A fact that may help to explain the *apparent* recent success of this model is that the period 1989–2000 happens to be a rare period in which predictions based on the climatology of the prior 25 yr are unusually good. This reflects the fact that the monsoon rainfall has been near normal every year during this period. Consequently, *any* forecast model that predicts near-normal rainfall during this period will have a relatively small mean-square error.

As with all statistical studies, we cannot eliminate the possibility that some set of predictors other than those used here could lead to significantly better predictions. Nor can we eliminate the possibility that analysis of longer datasets, or the use of nonlinear models, might produce better prediction models than those found here. Finally, our results cannot eliminate the possibility that dynamically based, coupled ocean–land–atmospheric models might predict monsoon rainfall with higher skill than linear models.

Acknowledgments. This research was supported by grants from NSF (ATM9814295), NOAA (NA96-GP0056), and NASA (NAG5-8202). The authors thank Ben Kirtman for many insightful comments that influenced this work.

REFERENCES

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel, 1994: *Time Series Analysis*. 3d ed. Prentice Hall, 598 pp.

Burnham, K. P., and D. R. Anderson, 1998: *Model Selection and Inference: A Practical Information Theoretic Approach*. Springer-Verlag, 353 pp.

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

Delecluse, P., M. Davey, Y. Kitamura, S. G. H. Philander, M. Suarez, and L. Bengtsson, 1998: Coupled general circulation modeling of the tropical Pacific. *J. Geophys. Res.,* **103,** 14 357–14 373.

Gowariker, V., V. Thapliyal, R. P. Sarker, G. S. Mandal, and D. R. Sikka, 1989: Parametric and power regression models: New approach to long range forecasting of monsoon rainfall in India. *Mausam,* **40,** 115–122.

——, ——, S. M. Kulshrestha, G. S. Mandal, N. Sen Roy, and D. R. Sikka, 1991: A power regression model for long range forecast of southwest monsoon rainfall over India. *Mausam,* **42,** 125–130.

Hastenrath, S., 1995: Recent advances in tropical climate prediction. *J. Climate,* **8,** 1519–1532.

Johnson, R. A., and D. W. Wichern, 1998: *Applied Multivariate Statistical Analysis*. 4th ed. Prentice Hall, 816 pp.

Kirtman, B. P., and J. Shukla, 2000: Influence of the Indian summer monsoon on ENSO. *Quart. J. Roy. Meteor. Soc.,* **126,** 213–239.

Lorenz, E. N., 1977: An experiment in nonlinear statistical weather forecasting. *Mon. Wea. Rev.,* **105,** 590–602.

Parthasarathy, B., A. A. Munot, and D. R. Kothawale, 1995: Monthly and seasonal rainfall series for all India, homogeneous regions and meteorological subdivisions: 1871–1994. Indian Institute of Tropical Meteorology Research Rep. RR-065, 113 pp. [Available from Indian Institute of Tropical Meteorology, Homi Bhabha Road, Pune 411008, India.]

Shukla, J., and D. A. Paolino, 1983: The Southern Oscillation and long-range forecasting of the summer monsoon rainfall over India. *Mon. Wea. Rev.,* **111,** 1830–1837.

——, and D. A. Mooley, 1987: Empirical prediction of summer monsoon rainfall over India. *Mon. Wea. Rev.,* **115,** 695–703.

——, ——, and D. A. Paolino, 1986: Long range forecasting of summer monsoon rainfall over India. *Ponteficiae Acad. Sci. Scripta Varia,* **69,** 147–178.

Sperber, K. R., and T. N. Palmer, 1996: Interannual tropcial rainfall variability in general circulation model simulations associated with the Atmospheric Model Intercomparison Project. *J. Climate,* **9,** 2727–2750.

Thapliyal, V., and S. M. Kulshrestha, 1992: Recent models for long range forecasting of southwest monsoon rainfall in India. *Mausam,* **43,** 239–248.

Trenberth, K. E., 1990: Recent observed interdecadal climate changes in the Northern Hemisphere. *Bull. Amer. Meteor. Soc.,* **71,** 988–993.

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 494 pp.

Webster, P. J., V. O. Magana, T. N. Palmer, J. Shukla, R. A. Tomas, M. Yanai, and T. Yasunari, 1998: Monsoons: Processes, predictability, and the prospects for prediction. *J. Geophys. Res.,* **103,** 14 451–14 510.