Artificial Skill due to Predictor Screening

TIMOTHY DELSOLE AND JAGADISH SHUKLA

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

(Manuscript received 7 January 2008, in final form 18 July 2008)

ABSTRACT

This paper shows that if predictors are selected preferentially because of their strong correlation with a prediction variable, then standard methods for validating prediction models derived from these predictors will be biased. This bias is demonstrated by screening *random numbers* and showing that regression models derived from these random numbers have apparent skill, in a cross-validation sense, even though the predictors cannot possibly have the slightest predictive usefulness. This result seemingly implies that random numbers can give useful predictions, since the sample being predicted is separate from the sample used to estimate the regression model. The resolution of this paradox is that, prior to cross validation, *all* of the data had been used to evaluate correlations for selecting predictors. This situation differs from real-time forecasts in that the future sample is not available for screening. These results clarify the fallacy in assuming that if a model performs well in cross-validation mode, then it will perform well in real-time forecasts. This bias appears to afflict several forecast schemes that have been proposed in the literature, including operational forecasts of Indian monsoon rainfall and number of Atlantic hurricanes. The cross-validated skill of these models probably would not be distinguishable from that of a no-skill model if prior screening were taken into account.

1. Introduction

A key issue in the construction of empirical forecast models is whether the model can make useful predictions of independent data—that is, data that were not used to construct the model. If a model predicts the available sample well but poorly predicts independent samples, then the model is said to have large *artificial skill*. It is generally recognized that two factors contribute to artificial skill: 1) complexity of the model and 2) number of models considered. The first factor is well appreciated by the seasonal forecasting community, as forecasters often adopt selection criteria that penalize complexity, such as cross validation, Akaike's Information Criterion, Bayesian Information Criterion, or Mallow's C_p (Barnston et al. 1994; DelSole and Shukla 2002; Kharin and Zwiers 2002). The second factor is perhaps less widely appreciated but no less important. The problem with considering a large number of models is that eventually one model will be found to fit the available data well, regardless, of its appropriateness.

A more subtle source of artificial skill is the method used to select predictors of a model. It is well established that the seasonal mean climate in a region may depend on sea surface temperatures (SSTs) in remote locations around the globe (Shukla and Kinter 2006). This fact implies that predictors for empirical seasonal forecasts may be large-scale fields requiring potentially thousands of variables to be specified. Choosing all possible relevant variables implies high model complexity and hence large artificial skill. To reduce artificial skill, the forecaster must be selective in the choice of predictors. Ideally, the forecaster selects predictors based on physical theory, perhaps guided by climate modeling experiments. Unfortunately, constraints derived from physical theory often do not reduce the pool of predictors to reasonable levels. Another approach is to approximate the predictors by a few parameters, such as their leading principal components, which capture maximum variance with the fewest components.

Another procedure for choosing predictors, which is the focus of the present paper, is *screening*. Screening is any procedure for choosing variables that preferentially includes or excludes certain characteristics of the *joint relation* between predictor and predictand. For instance, a forecaster may start with 1000 variables, but then

Corresponding author address: Timothy DelSole, 4041 Powder Mill Rd., Suite 302, Calverton, MD 20705-3106.
E-mail: delsole@cola.iges.org

DOI: 10.1175/2008JCLI2414.1

select the 10 most correlated variables with respect to the predictand. An important fact, quantified by Davis (1977) and Lanzante (1984), is that artificial skill arising from screening is larger than artificial skill arising from predictors chosen a priori (i.e., chosen by a nonscreening method). The degree of artificial skill increases with the number of predictors and pool of potential predictors and decreases with increasing sample size and true skill (Chelton 1983; Shapiro 1984; Shapiro and Chelton 1986).

The fact that screening also compounds artificial skill was pointed out by Michaelson (1987) in his classic paper on cross validation. Cross validation is the technique of setting aside a few observations, constructing a model from the remaining sample, and testing the model on the dataset aside, and then repeating this procedure by setting aside other observations in turn until all observations have been used exactly once for testing. Michaelson (1987) noted that, if predictors are selected by screening, both the screening and model building procedure must be cross validated. In this case, however, no single model is validated—the predictor set, and hence the prediction model, changes with each validation step; in essence, cross validation validates the procedure, not the model. Michaelson (1987) used cross validation to quantify the artificial skill in some model building procedures for predicting winter SSTs from fall SSTs.

Despite its dangers, not all sample-based procedures for selecting predictors are validated as a routine practice. For instance, a common procedure is to derive prediction models from the leading principal components of the data without validating the principal components themselves (i.e., the principal components are not recomputed in each validation step). This procedure is used, for example, to develop seasonal forecast models from linear inverse models (LIM; Penland and Magorian 1993), canonical correlation analysis (CCA; Barnston and Smith 1996), and constructed analogs (Van den Dool 2007). However, there exists a fundamental difference between selecting predictors based on screening and selecting predictors based on the leading principal components. Specifically, the former depends on the *joint* relation between predictor and predictand, whereas the latter does not. That is, the latter method selects principal components because they explain maximal variance, not because they are well correlated with the predictand. While principal components do in fact lead to biased estimates of *variance* in independent data (Lawley 1956), this bias is distinct and presumably uncoupled from artificial skill. Hence, selecting leading principal components of predictor variables, in decreasing order of variance, is not expected to cause the same bias as selecting predictors based on screening.

Although Davis (1977), Lanzante (1984), and Michaelson (1987) show that screening causes artificial skill, they do not show that, if screening is performed on all available data prior to validation, cross validation leads to biased estimates of skill. The purpose of this paper is to demonstrate this fact and to discuss the implications of this fact for certain types of forecasts.

The fundamental problem with validating a model after screening has been performed on all available data can be illustrated with a simple example. Consider predicting the variable y based on a predictor variable x which is chosen from a very large pool of variables x_1, x_2,. . . x_p, none of which really are related to y. While the majority of predictor variables will have no relation to y, a few predictor variables will have strong sample correlations with y, simply by virtue of a large pool of variables. Suppose a forecaster finds a predictor x that is *perfectly* correlated with y; that is, the sample correlation is one. A scatter diagram between x and y would reveal all points lying on a line. Next, consider applying leave-one-out cross validation to this dataset, in which the sample is partitioned into two subsets, one for constructing the model and the other for assessing the model. No matter how the data are partitioned, a least squares line fit based on any subset of points will perfectly predict the remaining points and thus appear to have perfect cross-validation skill. This example shows that if, prior to model construction, a predictor is selected because of its high correlation, cross validation is biased toward positive skill. The bias occurs because the withheld data that is ''left out'' in cross validation does not constitute an independent draw from the joint distribution of y and x, since, prior to model construction, *all* of the data had been used to find the x variable that covaried with y.

The sample correlation need not be exactly one for the above bias to occur—any nonzero correlation found from screening will lead to bias in cross validation. To demonstrate this bias, we perform a series of numerical experiments in which a linear regression model for predicting y is constructed from a set of predictors derived from *random numbers*. Since the predictors are random, we know that they cannot have the slightest predictive usefulness. Nevertheless, proceeding as if the random numbers are potentially useful predictors, we generate a set of random time series $x_1(t)$ $x_2(t)$. . . $x_P(t)$ and compute the correlation between each time series and $y(t)$. Even though $x_1(t)$ $x_2(t)$ . . . $x_P(t)$ are independent random time series, the sample correlation between some of these time series and $y(t)$ can be large, just by chance. We then select the time series with the largest absolute correlations, which we call the *screened predictors*. We next perform cross validation, in which some time steps of the screened predictors are withheld and a regression model is derived from the remaining data. We show that the model still gives skillful predictions of the withheld sample. Since one of the prediction variables is number of hurricanes, this result seemingly implies that random numbers can predict the number of hurricanes. The resolution of this paradox is that, prior to withholding the sample, *all* of the data had been used to find predictors that were well correlated with the prediction variable.

The bias caused by screening predictors appears to afflict several proposed forecasting schemes, including one by the authors (DelSole and Shukla 2002). In the latter study, the authors adopted predictors used in previous studies without questioning how the predictors were selected originally. Space limitations prevent us from documenting all forecasts in which screening is a potential cause of artificial skill. Consequently, we have chosen to focus on seasonal forecasts of Indian monsoon rainfall and the number of Atlantic hurricanes, as documented by Rajeevan et al. (2007) and Klotzbach and Gray (2004), respectively, because these forecasts are issued regularly and utilized by a large community. We show that the cross validated skill of these models are consistent with that of a no-skill model if screening were taken into account.

In the next section, we review particular forecast schemes in which screening plays a fundamental role. We describe our methodology for quantifying the effect of screening in section 3, and discuss the results in section 4. We conclude with a summary and discussion of our results.

2. Forecasts of Indian monsoon rainfall and number of hurricanes

We consider forecasts produced by the Colorado State University Tropical Meteorology Project (CSUTMP). Although CSUTMP predicts several hurricane-related quantities, the number of Atlantic hurricanes (NH) was chosen for analysis because of its wide interest. As discussed by Klotzbach and Gray (2004), the predictors of this model are derived from the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis by constructing composite maps of the August and September mean sea surface temperature, sea level pressure, and zonal velocity at 200, 850, and 1000 hPa, based on the 10 highest and 10 lowest years of Net Tropical Cyclone (NTC) activity. Areas with large correlations are boxed, provided the boxes span at least 10° longitude and 20° latitude. The field within the box is then averaged to construct a predictor time series. A similar procedure is applied to correlation maps between NTC and the same monthly mean fields and to correlation maps between NTC and forecast residuals based on one or two predictors. The resulting predictors are examined to ensure that the partial correlation of each predictor exceeds the 10% significance level, that the predictor is not strongly correlated with other predictors, and that the correlation with NTC does not drop too much if the data is subdivided.

We also consider forecasts of the Indian June–September Mean Rainfall (ISMR) produced by the Indian Meteorological Department (IMD). The method used by the IMD to choose predictors has not been documented—(Gowariker et al. 1989, 1991), Thapliyal and Kulshrestha (1992), and Rajeevan et al. (2007) merely state the predictors. We have received the following description about the IMD methodology (M. Rajeevan 2007, personal communication). First, correlation maps between ISMR and selected variables from the NCEP–NCAR reanalysis and SST data are computed. Then, boxes with correlations exceeding the 5% significance level are identified. The variable is averaged over the box to derive a time series. The resulting time series are then examined to ensure that ''running'' 21-yr correlations remain high, that a plausible physical connection exists between the time series and ISMR, that the difference in mean values between the largest and smallest ISMR values are statistically significant, and that the high correlations are not due to a few anomalous years.

In the following, we call both forecasts ''operational.'' In addition, we consider only forecasts issued in May or June, prior to the onset of the seasonal phenomena. Although forecasts by the IMD and CSUTMP are publicly available, the predictors used to construct the forecasts are not. Predictors of ISMR were provided by M. Rajeevan of the IMD, and predictors of number of Atlantic hurricanes were provided by P. Klotzbach of the CSUTMP.

Both operational forecasts share a common step: prior to model construction, a large dataset is searched for variables that are strongly related to the predictand. The precise details of the selection criteria differ between the two forecasts, but overall the criteria are designed to select variables that covary with the predictand. For instance, in both schemes, the predictor must maintain some fraction of its correlation when the data is subdivided. While the ultimate selection criteria may be more stringent than simply finding correlated time series, the extra stringency does not change the fact that screening is still being performed. The stronger the covariability, the more likely a variable will be selected. For sufficiently large pool of predictors, an arbitrarily large number of predictors will be selected, even if there is no real relation.

Since the precise model building procedures employed at the IMD and CSUTMP are under constant development, the above descriptions are likely to be obsolete by the time this paper is published. Indeed, while this paper was under review, Klotzbach (2007) developed a refined model building procedure using 1949–89 data for training and 1900–48 and 1990–2005 data for testing. We argue that the precise procedure is not important. Rather, the critical feature that compounds artificial skill is that predictors are selected by screening all data prior to model validation. The new scheme proposed by Klotzbach (2007) does not use all data for screening, and therefore may not be subject to the bias discussed here.

An important question is how many independent time series are examined in the above procedures. This number is not simply the number of grid points in a dataset, since neighboring grid points are correlated and hence do not constitute independent degrees of freedom. Stefanick (1981), North et al. (1982), and Bretherton et al. (1999) argue that a rational estimate of the effective number of degrees of freedom in a time varying field depends on the length scale L_1:

$$L_1 = \int_0^\infty \rho(r)\,dr, \tag{1}$$

where $\rho(r)$ is the spatial autocorrelation function (i.e., the correlation between a point and all points a distance r from it). This measure tends to underestimate the length scale of oscillatory autocorrelations, since the oscillations cancel in the integral. DelSole (2001) suggested an alternative measure in time series analysis, which in the spatial domain would be

$$L_2 = 2\int_0^\infty \rho^2(r)\,dr. \tag{2}$$

This measure avoids the cancellation problem and moreover is consistent with the L_1 measure for exponential autocorrelation functions. We use the L_2 measure in this paper.

We compute the effective number of spatial degrees of freedom by dividing the total area of the domain by the area of a circle with radius L_2. The spatial autocorrelation function for several April mean fields were estimated from the NCEP–NCAR reanalysis during the period 1948–2007; some examples are shown in Fig. 1. April averages were chosen because these fields are known immediately prior to monsoon or hurricane season. Since previous studies noted a difference in length scales between tropics and midlatitudes, the autocorrelations are computed separately for variables within 30° of the equator (the ‘‘tropics’’), and elsewhere (the ‘‘midlatitudes’’). Table 1 shows the corresponding length scales. We see that monthly mean wind variables tend to have short length scales (around 1200 km), while geopotential, temperature, and pressure tend to have large length scales (2000–4000 km). We have computed length scales from both formulas and found that $L_2 > L_1$ in all cases, so the use of (2), if anything, underestimates the number of degrees of freedom. If we add up all the degrees of freedom for all the variables examined by Klotzbach and Gray (2003), then the total number of degrees of freedom is 434. However, different variables within a local region may be correlated, so adding degrees of freedom may not be sensible. If only 1000-hPa zonal wind is considered, then the total number of degrees of freedom exceeds 200. Accordingly, we choose the conservative value of 200 and examine the implications of searching for predictors in a 200-variable dataset.

3. Methodology

We perform numerical experiments in which a linear regression model is constructed from random variables. The variable we want to predict, y_n, $n = 1, 2,\ldots, N$, is called the *predictand*, and the variables on which the prediction is based are called the *predictors*. Two distinct predictand data are considered: 1) observed June–September rainfall over India in each year (data provided by M. Rajeevan of the IMD), and 2) total number of Atlantic hurricanes in each year (downloaded from http://www.aoml.noaa.gov/hrd/hurdat/Data_Storm.html).

The predictors are generated by first drawing N random numbers independently from a Gaussian distribution with zero mean and unit variance. These numbers define a time series for a single predictor. This procedure is repeated P times. The resulting P vectors, each of length N, are called ‘‘predictors’’ even though they are random numbers. Next, the sample correlation between each predictor and predictand is computed, yielding P correlations. Then, the $S \leq P$ predictors with largest absolute correlation are selected. These predictors will be called *screened predictors*. This screening procedure differs from that used by CSUTMP and IMD in that our procedure relies exclusively on correlations without regard to spatial coherence or other properties of the time series. Nevertheless, these experiments are argued to be relevant

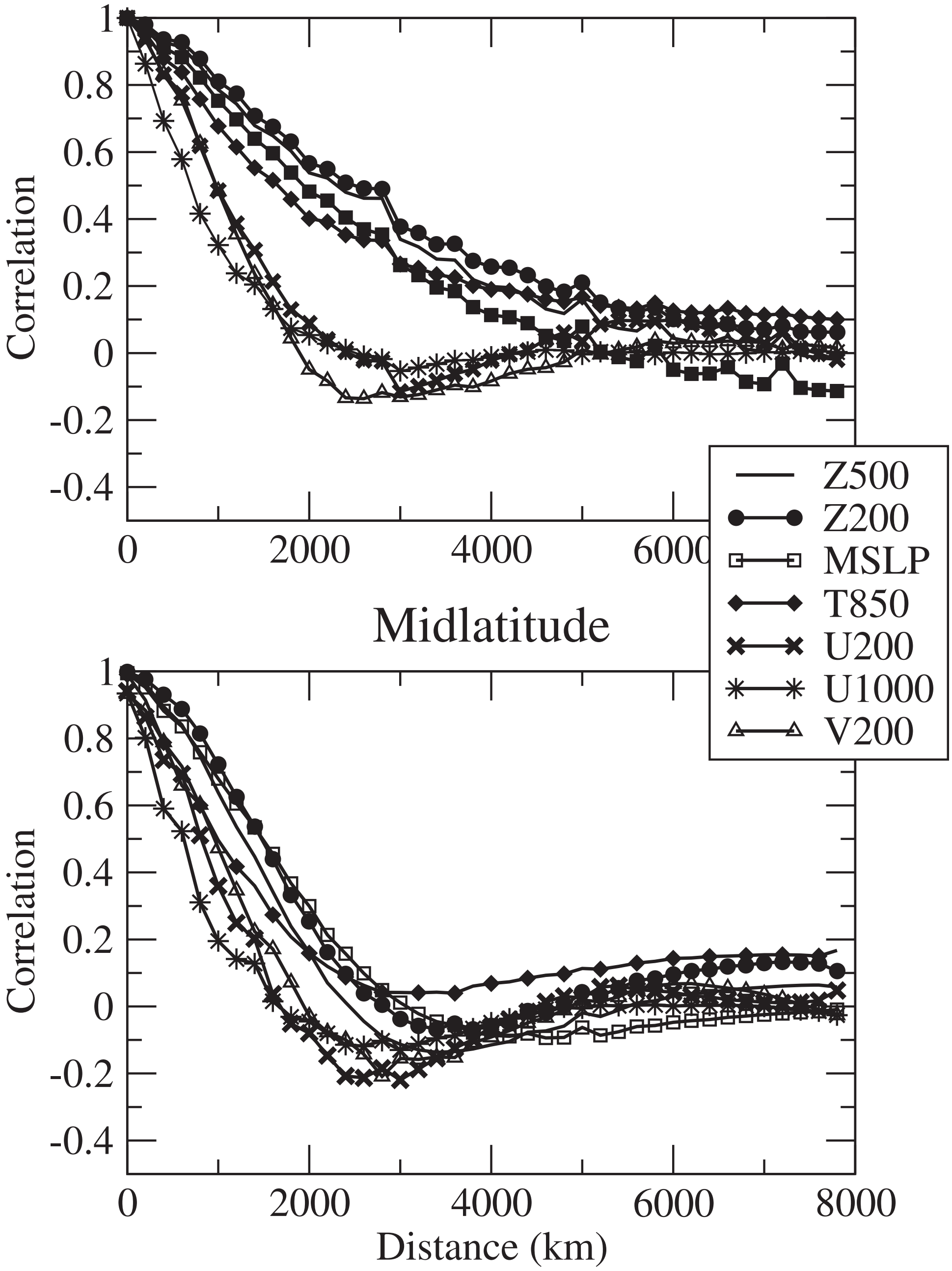

FIG. 1. Spatial autocorrelation of April mean fields estimated from the NCEP–NCAR reanalysis fields during the period 1948–2007. The fields are tabulated in Table 1 and denoted in the obvious way: (top) tropics and (bottom) midlatitudes.

because the precise screening procedure is not essential, rather it is the fact that screening is performed on the same dataset that is used to validate the model. Other screening procedures that approximate the operational screening procedures more closely will be considered at the end of the next section.

Having constructed random predictors, we next attempt to make a forecast of the predictand based on the random predictors. For this purpose, we assume, knowing full well that this is not the case, that the predictand and predictors are related linearly as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}, \tag{3}$$

where $\mathbf{y}$ is an N-dimensional vector specifying the predictand values, the rows of $\mathbf{X}$ give the predictor values

TABLE 1. Length scale and effective number of spatial degrees of freedom (dof) for April mean fields selected from the NCEP–NCAR reanalysis. The calculation of the length scale and degrees of freedom are described at the end of section 2.

	Midlatitude		Tropics		
Variable	Length (km)	Dof	Length (km)	Dof	Total dof
500-hPa geopotential height	1900	22	3800	6	28
200-hPa geopotential height	2200	17	4200	5	21
Mean sea level pressure	2100	18	2900	10	28
850-hP temperature	1600	32	2700	11	43
200-hPa zonal wind	1200	56	1500	36	92
1000-hPa zonal wind	800	127	1000	81	208
200-hPa meridional wind	1300	48	1500	36	84
Total dof		321		184	505

corresponding to each element of $\mathbf{y}$, $\boldsymbol{\beta}$ is a K-dimensional vector containing unknown regression parameters, and $\mathbf{w}$ is an N-dimensional random vector. The constant term is included by inserting a predictor whose value is always unity. Thus, except for a single column, all elements of $\mathbf{X}$ are random numbers. The least squares estimate of the regression parameters is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \tag{4}$$

where superscript T denotes the transpose operation. The least squares estimate of the predictand $\hat{\mathbf{y}}$, given the observed values of the predictors $\hat{\mathbf{X}}$, is

$$\hat{\mathbf{y}} = \hat{\mathbf{X}}\hat{\boldsymbol{\beta}}. \tag{5}$$

The final set of predictors used in the prediction model is determined by a *model selection* procedure. We consider only model selection procedures that use leave-one-out cross validation. In cross validation, one sample is withheld and the remaining samples are used to compute the least squares model. The resulting model then is used to predict the withheld sample. Repeating this procedure until each sample has been used exactly once as verification yields a set of forecast-verification pairs from which the sum square error can be computed. Importantly, the sum square error of leave-one-out cross validation can be computed without explicitly solving for the regression parameters in each stage of cross validation. The final result for the cross-validated sum square error (CVSSE) is given in Stone [1974, his (3.13)], which in our notation is

$$\mathrm{CVSSE} = \sum \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})_j^2}{[\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}]_{jj}^2}. \tag{6}$$

This formula reduces the number of calculations by more than an order of magnitude compared to explicit cross validation, which frees computational resources for more Monte Carlo samples. We typically report the square root of CVSSE, which we call the cross-validated (CV) standard error.

Two model selection procedures are considered in this paper: *all possible combinations* and *stepwise regression*. As the name suggests, the former method involves computing the cross-validated sum square error of all possible combinations of predictors and then choosing the combination that yields the smallest of these errors. In stepwise regression, we first choose the single predictor that minimizes the cross-validated sum square error and then subsequently add other predictors, one at a time, depending on whether it reduces the cross-validated sum square error. At each stage, the predictor that yields the greatest reduction in cross-validated sum square error is chosen. If no predictor in the set reduces the error, then the procedure halts and the resulting set of predictors are used to make forecasts. This form of stepwise regression differs from other standard forms in two ways: the criterion for accepting a predictor in the model is based on cross validation rather than an F-type significance test, and no predictor is removed from the model once it is selected, (i.e., we use ''forward selection'' with no ''backward removal''). This modified version of stepwise regression was chosen because we want to emphasize difficulties with cross validation, which may seem counterintuitive to some readers.

4. Results

Our first goal is to illustrate the bias due to searching large datasets, so the exact parameter choices for these experiments are not critical. Accordingly, we generate predictors by selecting the 10 random time series having largest absolute correlation with ISMR during the 30-yr period 1969–98, chosen from a pool of $P = 10$, 50, and 200 random time series. Figure 2 shows the cross-validated standard error of ISMR regression models formed from all possible subsets of the 10 screened predictors. The maximum and minimum correlations are given in parentheses in the figure. The top panel shows results for $P = 10$ and $S = 10$, which implies that no screening has been performed—10 random time series were generated and then selected. The model with minimum cross-validated standard error has 4 predictors. Since none of the models have genuine skill (because the predictors are random numbers), the 4-predictor model is the wrong choice. Note, however, that the minimum value is close to the error using 0 predictors and close to the standard error of ISMR. These results suggest a need to supplement the standard criterion in

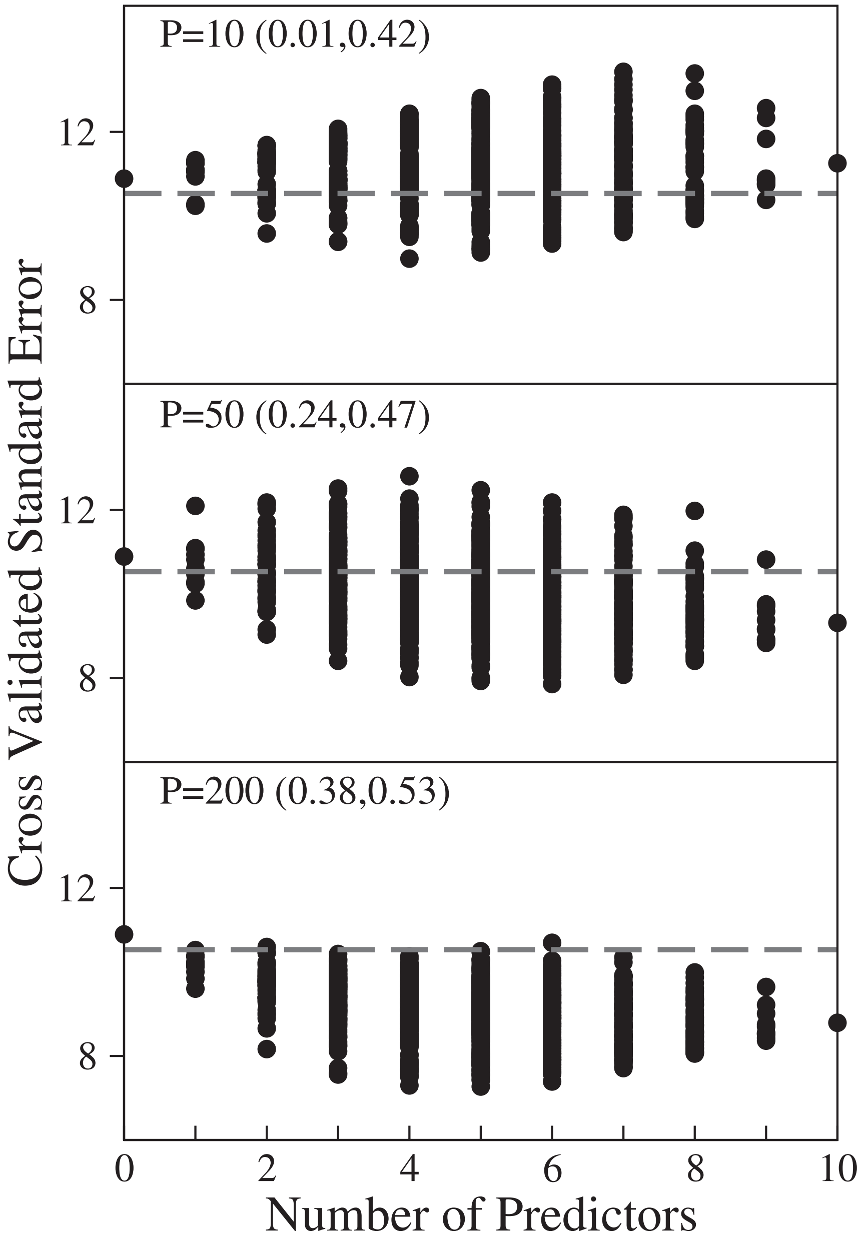

FIG. 2. Cross-validated standard error of regression models formed from all possible combinations of screened predictors. The screened predictors are the top 10 time series having the largest correlations with the predictand found from a total of *P random* time series, where the value of *P* is indicated in each panel: *P* (top) = 10, (middle) = 50, and (bottom) = 200. The predictand is the ISMR during the 30-yr period 1969–98. The number in parentheses in each panel gives the maximum and minimum correlation from the top 10 screened random time series. The dashed line in each panel indicates the standard deviation of the predictand during the period 1969–98.

cross validation by an additional criterion that checks whether the improvement due to adding a predictor is sufficiently large, a point also emphasized by DelSole (2007). The Monte Carlo technique used here may provide a quantitative basis for such a check.

Now consider a pool of predictors of size $P = 50$, but only the 10 time series with largest absolute correlations with ISMR are selected. The cross-validated error of all possible subsets of these $S = 10$ time series is shown in the middle panel of Fig. 2. First, the minimum

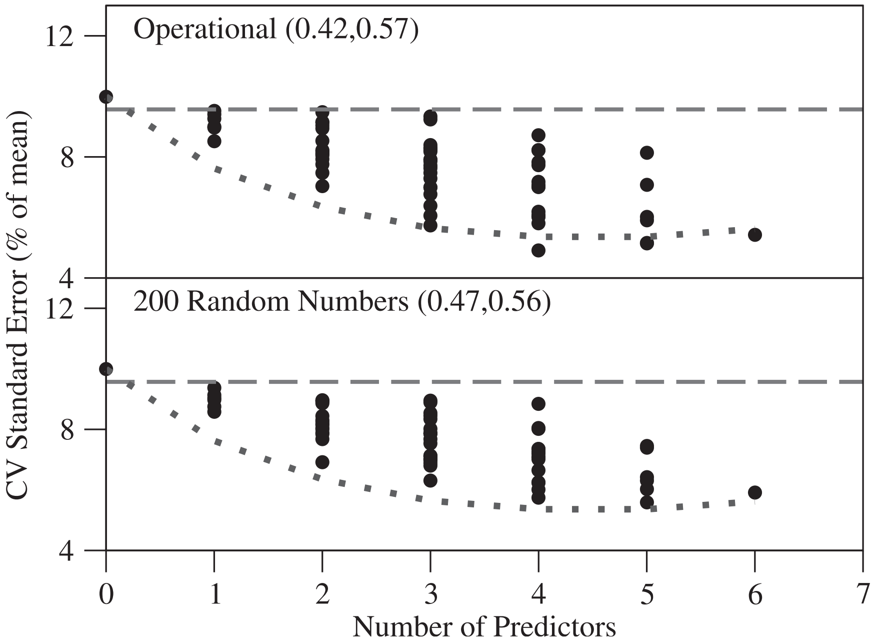

FIG. 3. (top) CV standard error of ISMR forecasts using the actual predictors of the May 2006 IMD operational forecast, and (bottom) a regression model derived from the 6 out of 200 random time series that are most correlated with ISMR. The numbers in parentheses give the minimum and maximum correlations between the six predictors and ISMR. The horizontal dashed line gives the standard deviation of ISMR during the period 1981–2004. The dotted curve gives the 5% significance threshold when screening is taken into account.

correlation of the screened predictors has risen by at least a factor of 10 relative to the case $P = 10$. This increase occurs in all experiments, though the exact factor is sample dependent. Second, the cross-validated error decreases with the number of predictors, reaching a minimum at 6 predictors. Increasing the original pool of predictors to $P = 200$ (bottom panel) leads to even lower cross-validated errors. This example shows that a model may have apparent skill in a cross-validation sense, even though the predictors cannot possibly have the slightest predictive usefulness.

It is instructive to compare the errors of forecasts based on random predictors with errors based on the actual predictors used in the operational forecast. At the time of this writing, the most recent forecast period was 2006. The IMD provided us with the six predictor time series they used to construct the May 2006 ISMR forecast. Similarly, the CSUTMP provided us with the six predictor time series they used to construct the April 2006 NH forecasts. In both cases, the forecasts are based on six predictors. Accordingly, we modify the above screening procedure to select at most six predictors from the large pool. We perform Monte Carlo simulations for the two predictands separately, using the appropriate time series length (the 2006 CSUTMP forecasts were trained on 1950–2005, while the 2006 IMD forecasts were trained on 1981–2005). The cross-validated errors of forecasts derived from the *actual predictors*, and based on a realization of the six most correlated time series out of $P = 200$ random time series, are shown in Figs. 3 and 4 . The dotted curve shows the 5% significance level for the minimum CV standard error as a function of the number of predictors, as determined by 1000 Monte Carlo simulations. We see that the errors lie almost completely above the dotted line, indicating that the no-skill hypothesis cannot be rejected based only on the observed degree of cross-validated skill. (If the 1% significance level were used, the dotted curve would extend below the panel boundary and all data points would lie above the significance curve, indicating lack of statistical significance.)

To construct the bottom panels of Figs. 3 and 4, only a few realizations of random time series were explored. Results for other realizations are surprisingly similar. This robustness is a consequence of random sampling from large datasets. To see this, let ρ_c be the $100\alpha\%$ significance level of a correlation derived from N independent, normally distributed random variables. By definition, then, the probability that a single random time series has a correlation exceeding ρ_c is α. However, the probability that at least one correlation out of P exceeds ρ_c is

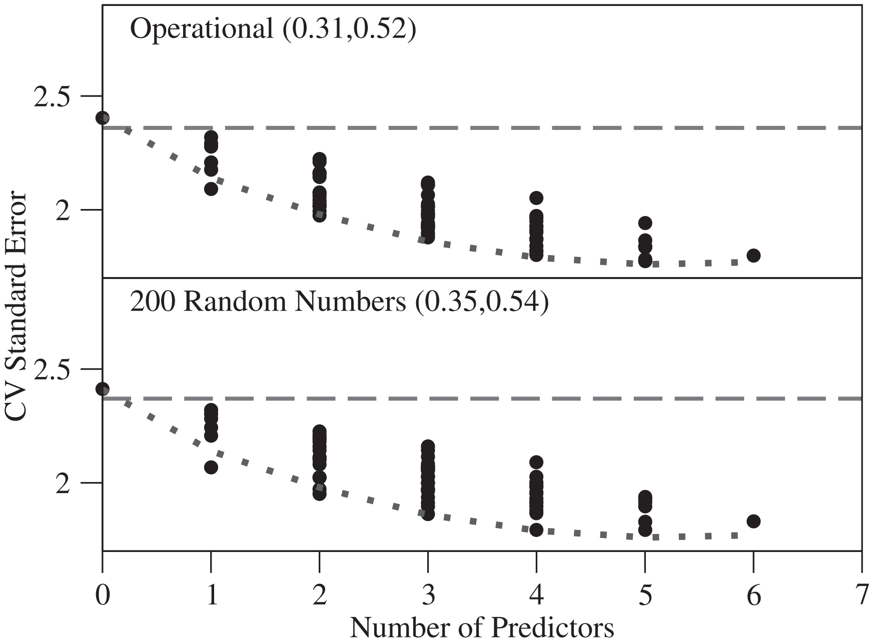

FIG. 4. (top) CV standard error of number of Atlantic hurricane (NH) forecasts using the actual predictors of the April 2006 CSUTMP operational forecast, and (bottom) a regression model derived from the 6 out of 200 random time series that are most correlated with NH. The numbers in parentheses give the minimum and maximum correlations between the six predictors and NH. The horizontal dashed line gives the standard deviation of NH during the period 1950–2001. The dotted curve gives the 5% significance threshold when screening is taken into account.

$$\text{prob} = 1 - (1 - \alpha)^P. \tag{9}$$

To put this formula in perspective, consider 100 time series, each of length 30 (as in the experiment on which Fig. 2 is based). The associated 5% significance threshold for the correlation coefficient is 0.34. Thus, about 5 out of 100 correlation coefficients are expected to exceed 0.34. However, the probability of finding at least one correlation coefficient exceeding 0.34 in a pool of 100 is $1 - (0.95)^{100} \approx 99.4\%$. Thus, a 5% rare event becomes a virtual certainty if just 100 time series are examined. More generally, if we search a large dataset, several time series with large correlations will inevitably be found. If we consider a new realization of time series, then while the actual numbers change, the upper range of correlations will be similar, in which case the best regression model derived from these time series will have similar skill.

Let us now consider results based on stepwise regression. Since this procedure is less demanding computationally, more realizations can be considered. Accordingly, we follow the previous procedure and generate P random time series of length $N = 30$, and then select the subset with the 10 largest absolute correlations with the predictand. We then repeat this procedure 1000 times, from which the median, upper and lower quartile, and 5% and 95% percentiles can be computed and displayed as ‘‘box-and-whisker’’ plots. The results are plotted in the bottom panels of Fig. 5. The cross-validated error decreases initially and reaches a minimum at 5 predictors, consistent with the results for all-possible-subsets regression. The top figure shows that the most frequent number of predictors chosen by stepwise regression is 5. Nevertheless, the selected models utilize random predictors and therefore cannot possibly have skill.

The results presented above do not prove that the skill of operational forecasts by the IMD and CSUTMP are consistent with a no-skill model because the screening procedures used in the different forecasts differ. Unfortunately, the different screening procedures have not been documented in sufficient detail to allow independent forecasters to discover the predictors exactly. For instance, the screening procedures involve a subjective component in which the physical plausibility of the predictor is assessed, the outcome of which differs from forecaster to forecaster. An alternative assessment is to measure the skill of real-time forecasts (i.e., forecasts issued before the verification becomes available). We show in Fig. 6 the standard error of three distinct forecasts over the period 1999–2007, the predictands of

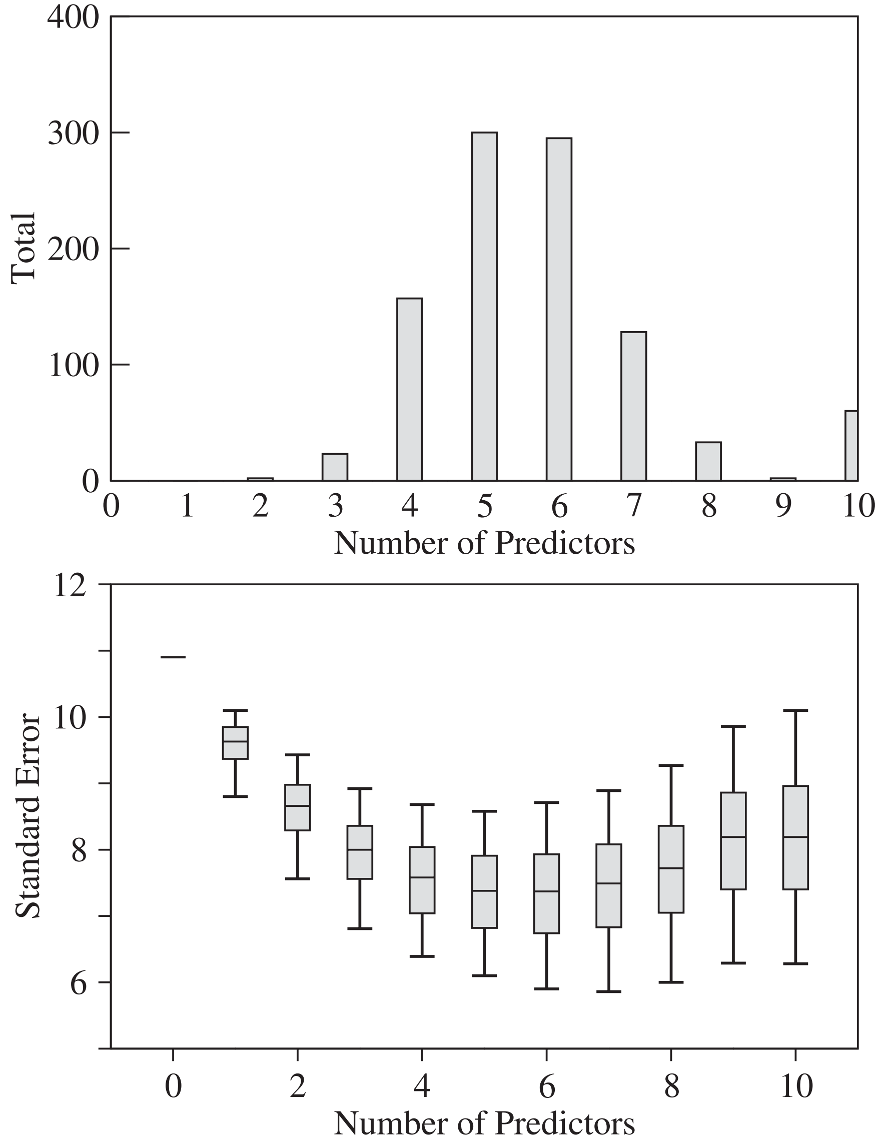

FIG. 5. Statistics of stepwise regression for predicting ISMR when applied to the top 10 out of 200 random time series having the largest absolute correlation with ISMR. (top) The number of cases in which stepwise regression chose a specific number of predictors; (bottom) box-and-whisker plots of the cross-validated standard error for each number of predictors. The box-and-whisker plots show the median as the centerline in the rectangle, the first and last quartile as the ends of the rectangle, and the 5% and 95% value as the ends of the error bars. The statistics were compiled from 1000 iterations.

which are tabulated in Tables 2 and 3: 1) a forecast based on the climatological means of the 15, 16, . . ., 30 yr preceding 1999, 2) the operational forecasts, and 3) forecasts by models with screened random predictors as selected by stepwise regression (without using the 1999–2007 data). Since the last forecasts are random we show box-and-whisker plots of the errors. The figure shows that the IMD forecasts are comparable to the lower quartile of the forecasts based on random predictors. Importantly, the IMD forecast does not perform better than forecasts based on the antecedent climatological means. The lack of real-time skill of the IMD models supports the hypothesis that the screening procedures used in operational forecasts have identified spurious relations.

In contrast to the IMD forecasts, the CSUTMP forecasts have significantly less mean square error than forecasts based on either the prior climatology or

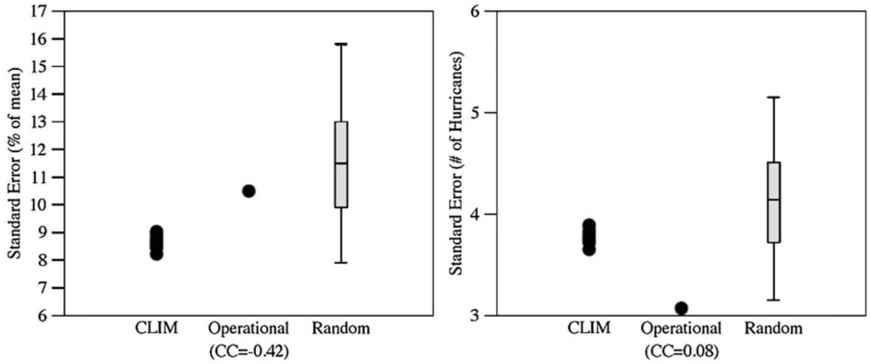

FIG. 6. (left) Standard error of three classes of forecast models applied to ISMR and (right) number of hurricanes during the period 1999–2007. ''CLIM'' refers to a forecast based on the climatology of the preceding data, from 15 to 30 yr; ''Operational'' refers to the. forecast issued by the IMD (left) and the CSUTMP (right); ''Random'' refers to the forecasts generated by the models with random predictors selected by the stepwise regression procedure, using the predictand only from the preceding 30 yr and whose statistics are illustrated in Fig. 5. The box-and-whisker plots show the median as the centerline in the rectangle, the first and last quartile as the ends of the rectangle, and the 5% and 95% value as the ends of the error bars.

random predictors. However, since the correlation skill during this period is small (i.e., 0.08), the reduced mean square error presumably arises from accurate predictions of the 1999–2007 mean. Specifically, the mean CSUTMP forecast during this period is 8.0, compared to the observed value of 7.7, whereas the observed mean in the 30 yr preceding 1999 is 5.5. The difference in means leads to poor mean square errors for the climatological forecast, but has no impact on correlation skill.

TABLE 2. Observed and forecasted Indian monsoon rainfall for the period June–September in units of percent of the 1941–90 mean (i.e., % of 89 cm). Also given are the corresponding mean, standard deviations, standard error of the forecast (stdv err), standard error of a forecast based on the prior 1970–95 mean (stdv clim), and the correlation between the forecast and observation (corr).

ISMR		
Year	Obs	Forecast
1999	−4.4	8
2000	−7.8	−1
2001	−9	−2
2002	−19.2	1
2003	2	−2
2004	−13	0
2005	−1	−2
2006	0	−7
2007	5	−5
Mean	−5.27	−1.11
Stdv	7.3	3.96
Stdv err		10.5
Stdv clim		8.8
Corr		−0.42

TABLE 3. Observed and forecasted number of Atlantic hurricanes for the period 1999–2007, as forecasted by the CSU Tropical Meteorological Project in late May–early June. Also given are the corresponding means, standard deviations, standard error of the forecast (stdv err), standard error of a forecast based on the 1969–98 mean (stdv clim), and the correlation between the forecast and observation (corr).

Number of hurricanes		
Year	Obs	Forecast
1999	8	9
2000	8	8
2001	9	7
2002	4	6
2003	7	8
2004	9	8
2005	15	8
2006	5	9
2007	4	9
Mean	7.67	8.00
Stdv	3.2	0.94
Stdv err		3.07
Stdv clim		3.75
Corr		0.08

To investigate the sensitivity of our results to screening criteria, we consider alternative screening procedures. In all cases, the first step is to order all *P* time series by decreasing absolute correlation with respect to the predictand. In the first procedure, we select the first *S* time series whose correlations exceed the 10% significance level; thus, this method may produce less than *S* time series, whereas previously we selected the time series with the leading *S* correlations, regardless of value. In the second procedure, we split the data in half and select the S time series whose correlations in both halves of the data maintain at least 90% of the original correlation. In the third, we split the data in half and select only the S time series whose correlations in both halves remain statistically significant at the 10% level with respect to the original time series. The results, shown in Fig. 7, are similar to each other and similar to the results shown Figs. 2–4, indicating that the bias is not sensitive to the precise screening procedure.

The sensitivity of the results to the total number of time series P, length of the time series N, and the number of screened predictors S is indicated in Table 4. In general, we find that the total number of predictors selected by stepwise regression is more strongly related to S than to N or P. Indeed, the number of screened predictors selected by stepwise regression is fairly well approximated by $S/2 \pm 2$, for the range of parameters examined.

5. Summary and discussion

Screening is the process of preferentially selecting a variable because of its strong covariability with the predictand. This paper demonstrates that if screening is not taken into account, then cross-validation methods overestimate forecast skill. This bias was illustrated by showing that if predictors are drawn from a large pool of *random numbers* by selecting only those that are strongly correlated with the predictand, then forecast models derived from the resulting predictors have substantial cross-validated skill. This result may seem surprising since, in cross validation, the data used to derive parameters in a regression model is separate from the data used to validate the model. Thus, without accounting for screening, this result seemingly implies that random numbers can provide useful forecasts of independent data. The resolution of this paradox is that, prior to cross validation, *all* the data had been used for screening. This bias does not afflict real-time forecasts since future data is not available for screening.

Several screening procedures have been proposed in the literature. A common part of these schemes is the identification of predictors from correlation maps. In the context of seasonal forecasts, the effective number of spatial degrees of freedom of monthly mean fields, as estimated from the spatial autocorrelation function, is about 30 for nonwind variables and 100–200 for wind variables. Thus, the total number of independent samples that are represented in the correlation maps is at least a few hundred. This result also implies that wind variables are more likely to be selected than nonwind variables, owing to their shorter length scales and hence their greater effective number in a global field [this may explain why 6 out of 7 predictors identified in Klotzbach and Gray (2003) are wind variables, and why 1000-mb zonal velocity, which has the smallest length scale in the set, is the single most prevalent predictor in the set].

This paper suggests that predictor screening has not been accounted for in the validation of some seasonal forecasts. We have focused specifically on the forecasts by the IMD and CSUTMP because they are operational and used by a large community, but we stress that many other schemes can be found in the literature that are problematic for the same reasons. In these forecasts, screening is not included in the cross validation, and the cross-validated errors are comparable to the errors expected from screening random numbers. Therefore, the skill of these models is consistent with a no-skill model after screening has been taken into account.

It is possible to argue that mean square error is not a good choice of skill measure, perhaps because it penalizes forecasts that have the correct correlation but incorrect amplitude. However, regression models specifically optimize mean square error, so it is appropriate to validate regression models using the same measure that they are designed to optimize. Nevertheless, alternative skill metrics also indicate no skill; for example, the correlation between observations and predictions over the past decade is negative for each set of forecasts (see Tables 2 and 3).

A characteristic feature of forecasts by the IMD and CSUTMP is that the predictors change with time. For instance, none of the predictors used by the IMD in 2007 were used by the IMD during 1988–2002. Similarly, only one predictor used by the CSUTMP in April 2007 was used in the April 1999 forecast model (namely, the February SST near the European coast). This apparent need to change predictors is an expected consequence of screening: screened predictors, being biased in the available sample, lose some or all predictive usefulness in independent data. It is sometimes argued that predictors must change with time because the climate system is nonstationary (Rajeevan et al. 2007). However, this hypothesis contradicts the basis of statistical prediction, namely that relations identified in the past will persist into the future. Forecasts based on nonscreening methods, such as those based on LIM or CCA, explicitly assume stationarity and seem to maintain skill as the models are updated with independent data. Also, if the system is nonstationary, then the relation between variables should not only degrade with time, but also should occasionally increase. To our knowledge, the predictive power of a screened predictor never increases in a statistically significant sense after the predictor has been defined.

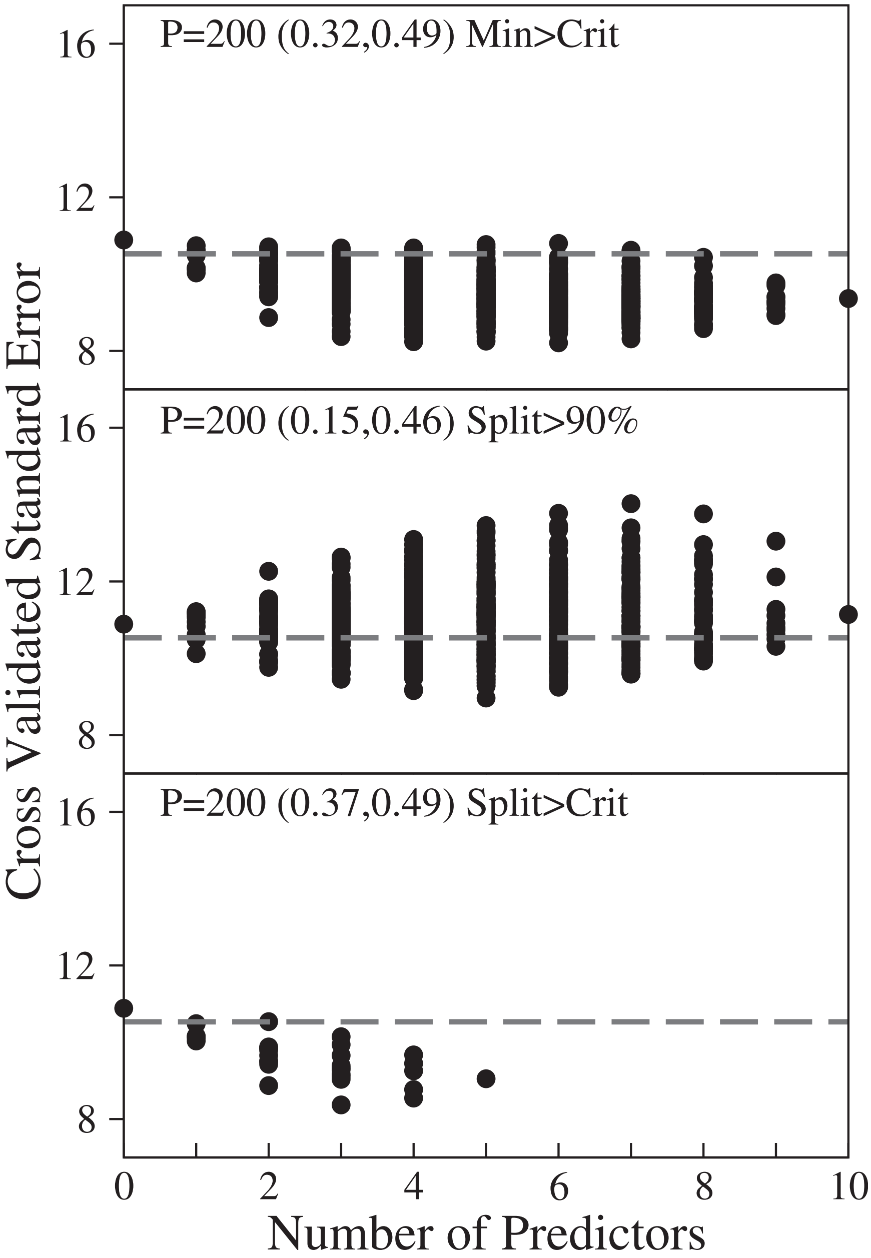

FIG. 7. Same as Fig. 2 for ISMR, but for three different screening procedures. In all cases, the time series that are maximally correlated with ISMR are selected, but with the following restrictions: (top) the correlations must exceed 10% significance, (middle) the correlation derived from either half of the data does not drop below 90% of the original correlation, or (bottom) the correlation derived from either half of the data is significant at the 10% level (based on the full time series). The last restriction is the most stringent and hence has fewer samples.

Jagannathan (1960) documents each regression model used by the IMD each year during the 67-yr period 1886–1960 (see his appendix 3). From this list, we find that the median time that a predictor remained in the IMD operational forecast is 10 yr. A natural question is how this interval compares with the interval required to decide that a screened predictor is useless for independent predictions. While several methods suggest themselves, a point of reference is the following. Consider repeatedly generating *N* independent random

TABLE 4. The most likely number of screened predictors chosen by stepwise regression, when the predictors are chosen from a pool of *P* random time series of length *N*, and the leading *S* time series with maximum absolute correlation with ISMR are selected for the screened predictors. The "most likely" number of predictors is the most frequently chosen number from 1000 Monte Carlo experiments.

N	*P*	*S*	Most likely number of predictors
26	100	5	3
26	500	5	5
26	100	10	5
26	200	10	5
26	500	10	5
26	1000	10	5
26	100	15	6
26	200	15	7
26	500	15	7
26	100	20	8
26	200	20	9
26	500	20	8
45	100	5	5
45	500	5	5
45	100	10	6
45	200	10	6
45	500	10	6
45	1000	10	7
45	100	15	8
45	200	15	8
45	500	15	8
45	100	20	9
45	200	20	10
45	500	20	10
45	500	30	14

pairs until the sample correlation exceeds the 5% significance level, and then monitoring the correlation as additional independent data are concatenated onto the original time series. This monitoring mimics a reasonable procedure a forecaster might follow after identifying a predictor. Monte Carlo simulations reveal that the average number of independent samples that can be added before the correlation drops below the original significance level is almost exactly a linear function of the original sample size *N*, with the average number being 7 for $N = 20$, and 11 for $N = 30$. Thus, the observed 10-yr interval for a predictor to remain in operational forecasts is consistent with the interval needed to reject a useless screened predictor identified originally in 20–30 samples.

Our conclusion that the cross-validated skill of operational models is consistent with a no-skill model, if screening is taken into account, implies that the operational forecasts should have no skill. This appears to be the case for the IMD forecasts: Montgomery (1940) showed that the forecasts by the IMD had no skill from the period 1921–35, and Gadgil et al. (2005) found that the correlation skill of the IMD forecasts during the period 1934–2004 was statistically insignificant. Earlier, Walker (1922) found that a regression equation derived in 1908 and used during the period 1909–21 had a correlation skill of 0.55, which is only marginally significant at the 5% level. It is interesting to note that Walker (1914) clearly recognized the problems due to screening and compensated for them by choosing a more stringent significance level.

The mean square error of CSUTMP forecasts since 1999 are substantially less than those based on climatology or random predictors. However, the correlation skill of these forecasts is small (i.e., 0.08), suggesting that the reduced mean square error arises primarily from accurately predicting the 1999–2007 mean. Ascertaining the reasons for why the regression models trained on past data accurately predicted that the 1999–2007 mean would be higher than the previous 30 yr is difficult because the specific forecast models used by CSUTMP change from year to year and the regression forecasts are subjectively adjusted before final issuance (P. Klotzbach 2007, personal communication). Nevertheless, our conclusion that the cross-validated skill itself is consistent with a no-skill model still holds.

Since screening compounds artificial skill, alternative methods of selecting predictors need to be developed. In the introduction, we mentioned some alternatives, including constraining the predictors based on physical theory or explained variance. Also, objectively defined screening procedures can be included in nested cross-validation procedures. Some novel methods for identifying predictive relations in large datasets have been developed by the statistics community, especially in data mining and machine learning (see Hastie et al. 2001), and their application to statistical climate prediction are worth investigating. It should be recognized, however, that artificial skill can intrude in subtle ways that even the most conscientious forecaster may fail to recognize. For instance, suppose a forecaster properly includes the screening procedure in cross validation and discovers that the prediction procedure is not skillful. The forecaster may then be tempted to modify the model construction procedure until the cross-validation procedure indicates skill. However, this approach constitutes screening in disguise, since, in effect, the "independent" samples are used to select the final model building procedure. The most satisfactory demonstration of skill is that based on real-time forecasts in which the predicted future event is completely unavailable at the time at which the forecast is issued.

Although this paper has focused on statistical prediction models, the results of this paper also have implications for dynamical prediction models. Specifically, the process of ''model development,'' which involves tuning and choosing parameterizations, can be interpreted as selecting one model out of a large set of models. If the selection criteria are based on how well the model predicts past events, model development effectively becomes a screening method and the bias discussed here becomes a significant problem with validating dynamical models.

Acknowledgments. We are grateful to P. Klotzbach and M. Rajeevan for providing numerous details of the forecasting methodology used at the CSUTMP and IMD, respectively, including providing data and preprints of papers. Dr. Klotzbach, in particular, identified an important error in our original manuscript. We also thank Michael Tippett and Ben Kirtman for numerous discussions regarding this work. An anonymous reviewer brought several relevant references to our attention and made several constructive comments that led to clarifications in the final manuscript. Andrew Colman, acting as reviewer, also made helpful comments that led to clarifications in the final manuscript. This research was supported by the National Science Foundation (ATM0332910), National Aeronautics and Space Administration (NNG04GG46G), and the National Oceanic and Atmospheric Administration (NA04OAR4310034).

REFERENCES

Barnston, A. G., and T. M. Smith, 1996: Specification and prediction of global surface temperature and precipitation from global SST using CCA. *J. Climate,* **9,** 2660–2696.

——, and Coauthors, 1994: Long-lead seasonal forecasts—Where do we stand? *Bull. Amer. Meteor. Soc.,* **75,** 2097–2114.

Bretherton, C. S., M. Widman, V. P. Dymnikov, J. M. Wallace, and I. Blade, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Climate,* **12,** 1990–2009.

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

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

DelSole, T., 2001: Optimally persistent patterns in time-varying fields. *J. Atmos. Sci.,* **58,** 1341–1356.

——, 2007: A Bayesian framework for multimodel regression. *J. Climate,* **20,** 2810–2826.

——, and J. Shukla, 2002: Linear prediction of Indian monsoon rainfall. *J. Climate,* **15,** 3645–3658.

Gadgil, S., M. Rajeevan, and R. Nanjndiah, 2005: Monsoon prediction–Why yet another failure? *Curr. Sci.,* **88,** 1389–1400.

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.

Hastie, T., R. Tibshirani, and J. Friedman, 2001: *The Elements of Statistical Learning.* Springer-Verlag, 552 pp.

Jagannathan, P., 1960: Seasonal forecasting in India: A review. FMU:1-80, India Meteorological Department, Pune, India, 120 pp.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate,* **15,** 793–799.

Klotzbach, P. J., 2007: Revised prediction of seasonal Atlantic basin tropical cyclone activity from 1 August. *Wea. Forecasting,* **22,** 937–949.

——, and W. M. Gray, 2003: Forecasting September Atlantic basin tropical cyclone activity. *Wea. Forecasting,* **18,** 1109–1128.

——, and ——, 2004: Updated 6–11-month prediction of Atlantic basin seasonal hurricane activity. *Wea. Forecasting,* **19,** 917–934.

Lanzante, J. R., 1984: Strategies for assessing skill and significance of screening regression models with emphasis on Monte Carlo techniques. *J. Climate Appl. Meteor.,* **23,** 1454–1458.

Lawley, N. D., 1956: Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika,* **43,** 128–136.

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

Montgomery, R. B., 1940: Report on the work of G. T. Walker. *Mon. Wea. Rev.,* **68** (Suppl. 39), 1–26.

North, G. R., F. J. Moeng, T. L. Bell, and R. F. Cahalan, 1982: The latitude dependence of the variance of zonally averaged quantities. *Mon. Wea. Rev.,* **110,** 319–326.

Penland, C., and T. Magorian, 1993: Prediction of Niño 3 sea surface temperatures using linear inverse modeling. *J. Climate,* **6,** 1067–1076.

Rajeevan, M., D. S. Pai, R. Anil Kumar, and B. Lal, 2007: New statistical models for long-range forecasting of southwest monsoon rainfall over India. *Climate Dyn.,* **28,** 813–828.

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.

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

Shukla, J., and J. L. Kinter III, 2006: Predictability of seasonal climate variations: A pedagogical review. *Predictability in Weather and Climate.* T. Palmer and R. Hagedorn, Eds., Cambridge University Press, 306–341.

Stefanick, M., 1981: Space and time scales of atmospheric variability. *J. Atmos. Sci.,* **38,** 988–1002.

Stone, M., 1974: Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc. Ser. A,* **36,** 111–147.

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

Van den Dool, H., 2007: *Empirical Methods in Short-Term Climate Prediction.* Oxford University Press, 215 pp.

Walker, G. T., 1914: Correlation in seasonal variation of weather III: On the criterion for the reality of relationship or periodicities. *Mem. India Meteor. Dep.,* **21,** 12–15.

——, 1922: Correlation in seasonal variation of weather VII: The local distribution of monsoon rainfall. *Mem. India Meteor. Dep.,* **23,** 23–39.