Empirical Probability Models to Predict Precipitation Levels over Puerto Rico Stations

NAZARIO D. RAMIREZ-BELTRAN

Department of Industrial Engineering, University of Puerto Rico, Mayagüez, Puerto Rico

WILLIAM K. M. LAU

Laboratory for Atmospheres, NASA Goddard Space Flight Center, Greenbelt, Maryland

AMOS WINTER

Department of Marine Sciences, University of Puerto Rico, Mayagüez, Puerto Rico

JOAN M. CASTRO

Department of Electrical and Computer Engineering, University of Puerto Rico, Mayagüez, Puerto Rico

NAZARIO RAMIREZ ESCALANTE

Department of Civil Engineering, University of Puerto Rico, Mayagüez, Puerto Rico

(Manuscript received 15 September 2005, in final form 27 April 2006)

ABSTRACT

A new algorithm is proposed to predict the level of rainfall (above normal, normal, and below normal) in Puerto Rico that relies on probability and empirical models. The algorithm includes a theoretical probability model in which parameters are expressed as regression equations containing observed meteorological variables. Six rainfall stations were used in this study to implement and assess the reliability of the models. The stations, located throughout Puerto Rico, have monthly records that extend back 101 yr. The maximum likelihood method is used to estimate the parameters of the empirical probability models. A variable selection (VS) algorithm identifies the minimum number of variables that maximize the correlation between predictors and a predictand. The VS algorithm is used to identify the initial point and the maximum likelihood is optimized by using the sequential quadratic programming algorithm. Ten years of cross validation were applied to the results from six stations. The proposed method outperforms both climatology and damped persistence models. Results suggest that the methodology implemented here can be used as a potential tool to predict the level of rainfall at any station located on a tropical island, assuming that at least 50 yr of monthly rainfall observations are available. Model analyses show that meteorological indices can be used to predict rainfall stages.

1. Introduction

Puerto Rico (PR), part of the Greater Antilles island chain, is located in the northeastern Caribbean Sea. Precipitation is primarily affected by troughs imbedded in easterly waves during summer months and cold fronts during winter months. The lowest amount of precipitation in PR occurs from December to March. The rainy season is characterized by two peaks, usually occurring during May and September.

A few researchers have attempted to simulate and predict the rainfall process in Puerto Rico. Comarazamy (2001) successfully used the regional atmospheric modeling system (RAMS) to simulate monthly rainfall over Puerto Rico during April 1998, which was an unusually wet month. However, RAMS failed to simulate the rainy season (i.e., August and September) because it was difficult to correct for initial soil moisture conditions.

Corresponding author address: Dr. Nazario D. Ramirez-Beltran, Department of Industrial Engineering, University of Puerto Rico, Mayagüez, 00680 Puerto Rico.
E-mail: nazario@ece.uprm.edu

DOI: 10.1175/MWR3290.1

© 2007 American Meteorological Society

Carter and Elsner (1997) used factor analysis to identify regions of mesoscale rainfall variance in nontropical storm convective rainfall (hurricanes and tropical storms were removed from the dataset). They analyzed 15 yr (1973–88) of data to regionalize Puerto Rico. They also extracted surface daily data during 3 yr (1977–79) and applied a partially adaptive classification tree to predict the occurrence of heavy rainfall events. Instead of predicting the exact rainfall amounts, they predicted the occurrence of light or heavy rainfall. Because their results were inconclusive, they pointed out that their statistical model should include satellite and radar products to improve tropical rainfall forecasts.

Carter (1999) also used factor analysis to identify convective rainfall regions in Puerto Rico that are statistically independent and describe similar rainfall variance. He conducted, in real time, an area-averaged 24-h quantitative precipitation forecast experiment over three 6-week periods in 1998. He found that his forecasts for heavy rain events were better than operational forecasts because of the absence of value-induced bias.

For this work, we plan to predict rainfall on a monthly basis using 101 yr of monthly information from 6 stations. The major forcing factors that modulate rainfall patterns in PR are identified by using a variable selection algorithm. The method is based on probability and empirical models. The parameters of the dynamic probability model are changing with time while the mathematical structure of the probability model remains unchanged. Parameters of the probability model are estimated at every point in time by using empirical functions that establish the relationship between a random vector that belongs to the probability model and a set of time series that are sequences of climatological observations. In time series literature, the empirical functions are known as lagged regression or transfer function models (Brockwell and Davis 2002; Box and Jenkins 1976; Pandit and Wu 1983; Wei 1990). A mathematical relationship between the dynamic probability model and the empirical functions is derived after taking the first moment of both the probability and the empirical models. Thus, the parameters of the dynamic probability model become a set of empirical functions.

The success of the probability model is highly dependent on the variable selection algorithm. If the appropriate variables are selected for an algorithm, then the probability model will provide reasonable probability forecasts. An algorithm for variable selection is introduced in this paper where the amount of predictors is larger than the number of observations. The algorithm is based on three regression concepts: parsimonious principle, stepwise selection, and multicollinearity problem. The principle of the algorithm is to divide the original set of predictors into smaller groups and to perform variable selections in each group. The variables that best explain the underlying predictand are selected by using the stepwise technique in such a way that the multicollinearity problem is avoided. The final estimates are obtained by applying the maximum likelihood (ML) method, which is used to estimate the parameters of the dynamic probability model. This is because ML estimates are asymptotically efficient, are invariant under linear transformations, and have a smaller mean-square error than other competing estimators (Bickel and Doksum 1977; Mood et al. 1974). The resulting estimation task consists of solving a constrained nonlinear optimization problem consisting of two parts. The first step is to select an appropriate initial point. The second step is to apply the sequential quadratic programming (SQP) algorithm to solve the constrained nonlinear optimization problem (Reklaitis et al. 1983; Bazaraa et al. 1993). If an initial point is carefully selected, then the nonlinear algorithm will converge to a satisfactory local maximum to assure that the appropriate parameters of the probability model are obtained. The dynamic probability model and the empirical regression equations are used to compute the probability that at a given station and a particular month, the rainfall level will be in one of the following stages: exceeds normal, equals normal, and below normal. Following a similar idea of Carter and Elsner's (1997), instead of predicting the exact rainfall amounts, we are predicting the occurrence of rainfall levels. Predicting the exact amount of rainfall at each station is extremely difficult since the amount of rainfall depends of the interaction of local thermodynamical information, which is not available for centenary records.

The observed data and meteorological indices are described in section 2. The dynamic probability model is introduced in section 3, where the variable selection algorithm and empirical models are also described. Section 4 presents a numerical example describing the major details of the prediction scheme. Section 5 summarizes the cross-validation results, and the conclusions are presented in section 6.

2. Data

The precipitation data used for this study come from six Puerto Rico weather stations: Coloso, Isabela, Manati, Maunabo, Mayagüez, and San Juan (Fig. 1). The station records were provided by the National Oceanic and Atmospheric Administration–Cooperative Institute for Research in Environmental Sciences (NOAA–CIRES) Climate Diagnostics Center and by

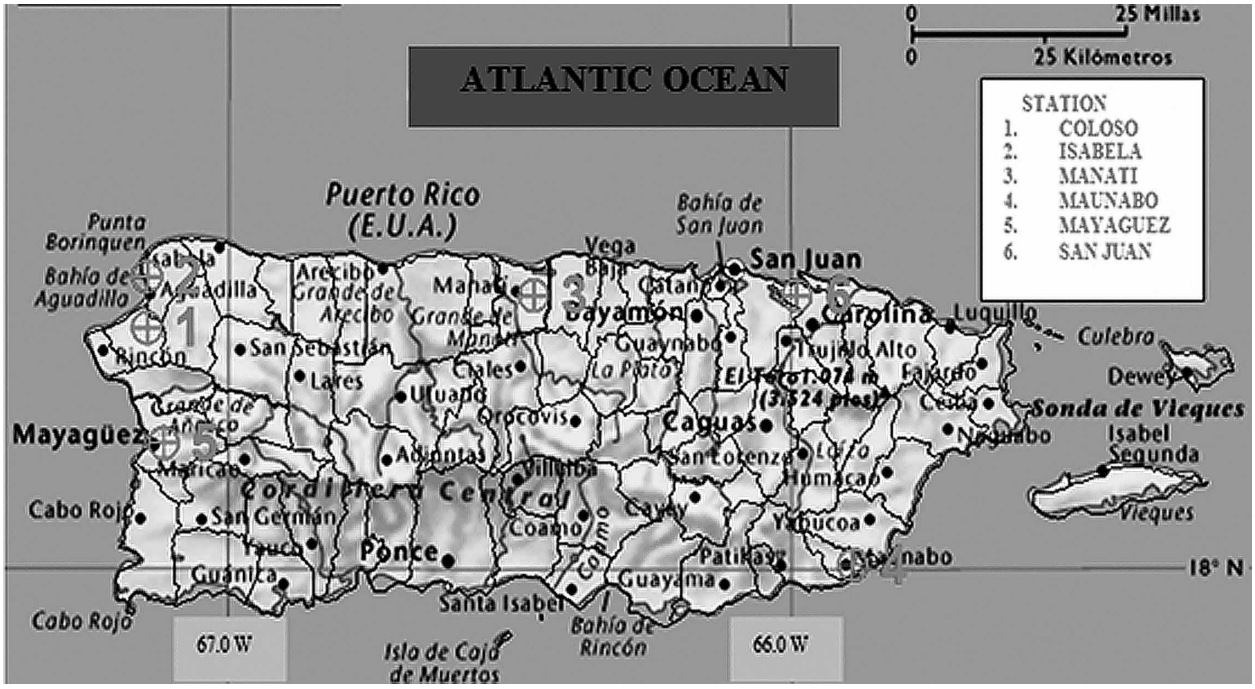

FIG. 1. Map of PR pointing out the location of the six stations. The numbers included on the map indicate the station locations and the names of the stations are listed on the upper right-hand side.

the Puerto Rico Climatology Office. The weather station records are from 1901 to 2001 and are resolved into monthly observations organized into 12 variables. In addition, five well-known meteorological indices (Table 1) were chosen as possible predictors of Puerto Rican precipitation variability: North Atlantic Oscillation (NAO; Hurrell 1995), Artic Oscillation (AO; Thompson and Wallace 1998), Southern Oscillation index (SOI; Trenberth 1984), Sahel Rainfall index (SRI; Janowiak 1988), and Brazil Rainfall index (BRI; Ward and Folland 1991). Gray et al. (2005) have shown that NAO, ENSO, AO, and SRI are correlated with the climate conditions in the North Atlantic basin and that the upper air data associated with the northeastern Brazilian region exhibit superior hindcast prediction skill over previous models. Taylor et al. (2002) show that interannual variability in the early season is influenced strongly by anomalies in the sea surface temperatures of the tropical North Atlantic, with positive anomalies over a narrow latitudinal band (0°–20°N) being associated with enhanced Caribbean rainfall. Also, a study by Malmgren et al. (1998) showed a close relation between NAO and PR rainfall over a period of nearly 100 yr.

3. Methodology

The proposed prediction scheme was divided into monthly bins and into three precipitation categories: above normal (>90th percentile), equal to normal (between 10th and 90th percentile), and below normal (<10th percentile). These stages are given the terms excess, normal, and scarce, respectively. A graphical display of the prediction technique is shown in Fig. 2.

TABLE 1. Meteorological variables. The first column indicates the number of the variable and the second column indicates the variables used in this work. The first 12 variables represent local information from the stations and the last 5 variables are meteorological indices that were used to correlate with the rainfall process at each Puerto Rican station.

No.	Variable name
1	Monthly rainfall at Coloso
2	Monthly rainfall at Isabela
3	Monthly rainfall at Manati
4	Monthly rainfall at Maunabo
5	Monthly rainfall at Mayagüez
6	Monthly rainfall at San Juan
7	Monthly average air temperature at Coloso
8	Monthly average air temperature at Isabela
9	Monthly average air temperature at Manati
10	Monthly average air temperature at Maunabo
11	Monthly average air temperature at Mayagüez
12	Monthly average air temperature at San Juan
13	North Atlantic Oscillation
14	Arctic Oscillation index
15	Sahel rainfall index
16	Northeast Brazil rainfall index
17	Southern Oscillation index

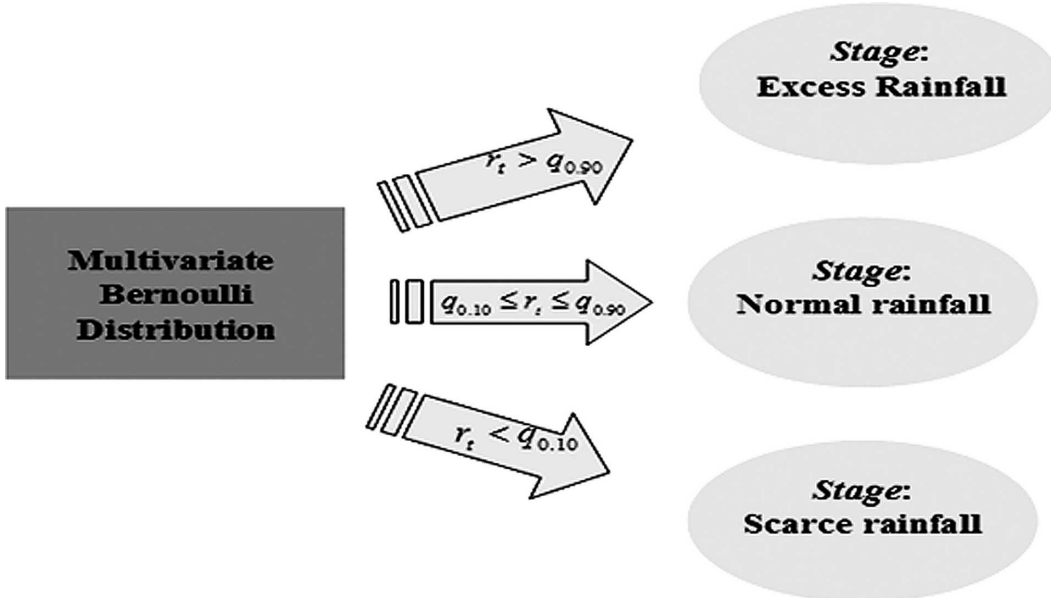

FIG. 2. The general prediction scheme. The MB distribution is used to identify the probability of rainfall stage, which would be excess, normal, or scarce.

a. Predicting a rainfall stage

Each of the stages is a mutually exclusive event and therefore the rainfall processes can be represented by a stochastic sequence. The stage of a single month for a particular year can be modeled by the multivariate Bernoulli distribution.

A random vector at time t, $\mathbf{Y}_t$, has the multivariate Bernoulli distribution if its probability mass function can be written as follows:

$$f(y_{1,t}, y_{2,t}, \ldots, y_{u,t}) = p_{1,t}^{y_{1,t}} \quad p_{2,t}^{y_{2,t}} \cdots p_{u,t}^{y_{u,t}}, \quad t = 1, \ldots, n, \tag{1}$$

$$\sum_{k=1}^{u} y_{k,t} = 1, \quad \sum_{k=1}^{u} p_{k,t} = 1, \quad t = 1, \ldots, n, \tag{2}$$

$$y_{k,t} = 0, 1 \quad \text{and} \quad p_{k,t} \ge 0, \quad k = 1, \ldots u \quad \text{and} \quad t = 1, \ldots, n, \tag{3}$$

where u is the number of stages, n is the number of available observations, and $p_{k,t}$ is the probability of success that the random variable $y_{k,t}$ is at the kth stage at time t.

Since at any given month the rainfall process is defined by three mutually exclusive stages, its probability mass function can be written as follows:

$$f(y_{1,t}, y_{2,t}, y_{3,t}) = p_{1,t}^{y_{1,t}} \quad p_{2,t}^{y_{2,t}} \quad p_{3,t}^{y_{3,t}} \quad t = 1, \ldots, n \tag{4}$$

$$\begin{aligned} 1 &= y_{1,t} + y_{2,t} + y_{3,t}, \\ 1 &= p_{1,t} + p_{2,t} + p_{3,t}, \\ y_{k,t} &= 0, 1 \quad p_{k,t} \ge 0 \quad \text{for} \quad k = 1, 2, 3 \\ t &= 1, \ldots, n, \end{aligned} \tag{5}$$

where

$$y_{1,t} = \begin{cases} 1, & \text{if} \quad r_{i,t} > q_{i,2} \\ 0, & \text{otherwise} \end{cases} \quad y_{2,t} = \begin{cases} 1, & \text{if} \quad r_{i,t} < q_{i,1} \\ 0, & \text{otherwise} \end{cases}$$

$$y_{3,t} = \begin{cases} 1, & \text{if} \quad q_{i,1} \le r_{i,t} \le q_{i,2} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The variable $r_{i,t}$ is the amount of rainfall for the ith station at month t, $q_{i,1}$ is the rainfall 10th percentile for the ith station, $q_{i,2}$ is the rainfall 90th percentile for the ith station, and $p_{k,t}$ are the parameters of the multivariate Bernoulli (MB) distribution and each one represents the probability that the rainfall in the tth month is in the kth stage. In this work, it is assumed that 10% of the rainfall events are in excess, 80% are normal, and 10% are scarce.

The parameters of the MB distribution will change through time, and consequently the parameters of the distribution could be expressed by climatological variables, which also change through time. Therefore, the dynamics of the probability model are captured by ob-

serving climatological variables. Empirical models are needed to express the relationship between the parameters of the probability model and the observed climatological variables. Probabilistic forecast has been traditionally based on logistic regression (Chiu and Kedem 1990; Gahrs et al. 2003; Wilson 2004). Recently, Sohn et al. (2005) compared four methods (linear regression, logistic regression, neural networks, and decision tree) to predict the occurrence of heavy rainfall events in Korea. They found that logistic regression provides the best validation results. Gahrs et al. (2003) also found that logistic regression is significantly more skillful than linear regression. Based on these results, the following empirical models were postulated to estimate the parameters of the dynamic probability model:

$$y_{k,t} = \frac{1}{1 + e^{-\gamma_{k,t}}} + \varepsilon_{k,t}, \quad k = 1, 2, 3 \quad t = 1, \ldots, n, \tag{7}$$

where

$$\gamma_{k,t} = a_{0,t,k} + \sum_{(s_j,t,g_j)\in A_k} a_{j,t,k} z_{s,t-g_j} + \varepsilon_{k,t},$$
$$k = 1, 2, 3, \quad t = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, \text{np},$$
$$g_j = 1, 2, \ldots, 24, \quad \text{and} \quad s_j = 1, 2, \ldots, \text{nv}, \tag{8}$$

where $z_{s_j,t-g_j}$ is the sth meteorological variable with a subscript that belongs to set A_k. The selected meteorological variables are listed in Table 1. Here, A_k is a set of subscripts for the kth stage defined as follows: s_j represents the variable name, t indicates the month, g_j is the time lag in months, nv is the total number of variables, np is the number of predictors in a specific equation, and n is the number of available observations. In this study, there is no physical argument to justify the lags used in the model. However, a high correlation identified in a predictor with a particular lag is an indication that there is a probability that the underlying lag has a physical mechanism, which may need further studying. The elements of the set A_k correspond to the variables that significantly contribute to explain the variability of $y_{k,t}$, which is a sequence of Bernoulli variables defined by Eq. (6). The constant $a_{j,t,k}$ is the jth regression coefficient at the ith month and at the kth stage, and $\varepsilon_{k,t}$ is a sequence of independent random variables with mean zero and constant variance. The success of the proposed model relies on the identification of meteorological variables that are highly correlated with the behavior of the Bernoulli variables. A variable selection procedure is proposed in this study and is presented at the end of this section.

Model 4 is a probability representation of a rainfall stage for a given month, and model 7 is an empirical representation of the same stage. Therefore, it is expedient to express a relationship between the empirical and the probability models to derive a parameter relationship between each model. This relationship is obtained by determining the first moment of both empirical and probability models. The first moment of the probability model is obtained as follows:

$$\begin{aligned} E(\mathbf{Y}_t) &= \sum \mathbf{Y}_t f(\mathbf{Y}_t) \\ &= [1 \quad 0 \quad 0] f(1, 0, 0) + [0 \quad 1 \quad 0] f(0, 1, 0) \\ &\quad + [0 \quad 0 \quad 1] f(0, 0, 1) \\ &= [p_{1,t} \quad p_{2,t} \quad p_{3,t}], \end{aligned}$$

where E is the mathematical expectation operator. Thus, it follows that

$$E(y_{k,t}) = p_{k,t} \quad k = 1, 2, 3, \quad t = 1, \ldots, n. \tag{9}$$

The expected value of the random variable $y_{k,t}$ given a set of meteorological variables (z) can be written as follows:

$$E(y_{k,t}|z) = \frac{1}{1 + e^{-\gamma_{k,t}}}, \quad k = 1, 2, 3, \quad t = 1, \ldots, n. \tag{10}$$

Therefore, assuming that historical information of meteorological variables is known up to time t, the relationship between the empirical model and the probability model can be expressed as follows:

$$p_{k,t} = \frac{1}{1 + e^{-\gamma_{k,t}}}, \quad k = 1, 2, 3, \quad t = 1, \ldots, n. \tag{11}$$

The parameters of the dynamic probability model are estimated by using the maximum likelihood method. The maximum likelihood function for the MB distribution is as follows:

$$L(\mathbf{p}) = \prod_{t=1}^{n} p_{1,t}^{y_{1,t}} \quad p_{2,t}^{y_{2,t}} \quad p_{3,t}^{y_{3,t}} \tag{12}$$

$$1 = y_{1,t} + y_{2,t} + y_{3,t}, \quad y_{k,t} = 0, 1,$$
$$1 = p_{1,t} + p_{2,t} + p_{3,t}, \quad \text{and} \quad p_{k,t} \geq 0 \quad \text{for}$$
$$k = 1, 2, 3, \quad t = 1, \ldots, n. \tag{13}$$

Maximizing the function [Eq. (12)] is equivalent to maximizing its natural log. Using Eq. (11), the maximum likelihood estimators for a can be obtained after maximizing the following expression:

$$J(\hat{\mathbf{a}}) = \sum_{t=1}^{n} \left[y_{1,t} \ln\left(\frac{1}{1+e^{-\hat{\gamma}_{1,t}}}\right) + y_{2,t} \ln\left(\frac{1}{1+e^{-\hat{\gamma}_{2,t}}}\right) + y_{3,t} \ln\left(\frac{1}{1+e^{-\hat{\gamma}_{3,t}}}\right)\right] \tag{14}$$

subject to

$$\frac{1}{1+e^{-\hat{\gamma}_{1,t}}} + \frac{1}{1+e^{-\hat{\gamma}_{2,t}}} + \frac{1}{1+e^{-\hat{\gamma}_{3,t}}} = 1, \tag{15}$$

where

$$\hat{\gamma}_{k,t} = \hat{a}_{0,t,k} + \sum_{(s_j,t,g_j)} \hat{a}_{j,t,k} z_{s_j,t-g_j}. \tag{16}$$

The variables $\hat{\gamma}_{k,t}$ and $\hat{a}_{j,t,k}$ are the estimates of $\gamma_{k,t}$ and $a_{j,t,k}$, respectively.

b. Parameter estimation

The constrained nonlinear optimization problem was solved by using a two-step strategy. The first step consists of applying a variable selection algorithm to identify the predictors that best correlate with the state variables, $y_{k,t}$. The best predictors are used to generate the appropriate initial point that guarantees that the nonlinear optimization algorithm will converge to a satisfactory local maximum. The first step is always the key in deriving a reasonable solution and consists of obtaining a unique initial point, which is derived by fitting a linear regression model. The second step consists of using a nonlinear optimization algorithm to estimate the parameters of the MB distribution.

1) Variable selection algorithm

A suitable requirement for an empirical model is that it (i) accomplishes the parsimonious principle, (ii) maximizes the explained variability, and (iii) does not exhibit the multicollinearity problem. The parsimonious principle consists of selecting the smallest number of predictors that maximize the correlation between the predictors and predictand. The multicollinearity problem occurs when the predictors are approximately linearly dependent, resulting in large increments in the variance of the predictors. It should be noted that the multicollinearity problem will produce extremely poor regression coefficients and consequently misleading predictions even though R^2 is close to one, where R^2 is the proportion of variability explained by the regression model and is usually known as the coefficient of multiple determination (Montgomery et al. 2001). Thus, if the multicollinearity problem is present, it must be removed before a prediction is computed. Removing this problem eliminates redundant information, consequently decreasing the variance of the prediction (i.e., only efficient estimators will be selected).

The variable selection (VS) algorithm, which is introduced in this study, selects a model that meets the previous characteristics. This algorithm is general, although it is especially useful for the cases where there are more predictors than observations. The VS algorithm includes three major steps.

(i) Step 1

The first step consists of designing the number of subgroups that will be created in a given dataset. The number of subgroups should be selected in such a way that the number of predictors included in each group provides enough degrees of freedom to properly estimate the regression coefficients in each group. The number of predictors in each group is called the group size. An empirical rule consists of selecting the integer number of the following ratio:

$$\mathrm{gs} \leq \frac{n}{3}, \tag{17}$$

where gs is the group size and n is the total number of observations. It should be noted that the degree of freedom, df, associated with errors in the regression model is $\mathrm{df} \geq (2n - 3)/3$.

The group size is often larger than the required size; however, the group size will also be controlled by the multicollinearity rule (step 3).

(ii) Step 2

The second step uses the stepwise algorithm to select the best predictors from each group (Montgomery et al. 2001). The mean square error (MSE) for each group is computed and the group that exhibits the minimum MSE is selected. The variables from a group that provides the minimum MSE ($\mathrm{MSE}_{\mathrm{min}}$) will be called the winner variables. The best variables from each group that does not contain the winner set are studied to determine whether or not they can improve the winner variable set. The winner variables will join the best predictors from a no-winner set (one at a time), and the stepwise procedure is used again to determine whether or not the MSE associated with these variables is smaller than the $\mathrm{MSE}_{\mathrm{min}}$. If that is the case, the $\mathrm{MSE}_{\mathrm{min}}$ is replaced by the new MSE and the winner variable set is also updated. This process is repeated over and over until all no-winner variable sets have been tested.

(iii) Step 3

The third step consists of testing whether or not the multicollinearity problem is present in the winner

dataset. Compute the eigenvalues of the $\mathbf{Z}$ matrix, where $\mathbf{Z} = \mathbf{z}'\mathbf{z}$ and $\mathbf{z}$ is a matrix whose columns are the predictors defined in Eq. (8). Compute the index number, ixn, as follows:

$$\text{ixn} = \frac{\lambda_{\max}}{\lambda_{\min}}, \tag{18}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and the minimum eigenvalues of the $\mathbf{Z}$ matrix, respectively. An empirical rule indicates that the multicollinearity problem is present when ixn > 100 (Montgomery et al. 2001).

The multicollinearity problem can be solved by using ridge regression, principal components, or by removing redundant predictors. The last technique was implemented in this research to conform to the parsimonious principle and it can be described as follows. Let $\boldsymbol{\xi}$ be the eigenvector associated with the minimum eigenvalue, $\lambda_{\min}$. It can be stated that the dimension of the $\boldsymbol{\xi}$ vector corresponds to the number of predictors included in the regression model. The locations of the predictors in the regression model are associated with the $\boldsymbol{\xi}$ elements. Thus, the first element of $\boldsymbol{\xi}$ is associated with the first predictor of the regression model, the second element of $\boldsymbol{\xi}$ is associated with the second predictor of the regression model, and so on. The predictor that will be removed is the one that corresponds to $\boldsymbol{\xi}_{\max}$, where $\boldsymbol{\xi}_{\max}$ is the maximum absolute value of the elements of $\boldsymbol{\xi}$. The variables that do not exhibit the multicollinearity problem will be the final winner variables for the underlying group. The winner variables provide the elements of the initial point for the nonlinear optimization routine.

The initial point is a vector that includes the regression coefficients, a, from three stages (excess, normal, and scarce). It should be mentioned that the initial point selected by the VS algorithm is a unique point and consequently causes the optimization routine to converge at the same local maximum, assuming that the predictors are physically related to the predictand.

If the underlying regression equation is linear, no additional work is needed to create the initial point. However, in this case, the regression Eq. (7) is not linear and consequently an additional computation is required since the stepwise algorithm works only with linear regression equations. The strategy proposed here is to use the observed rainfall in each month to estimate $p_{k,t}$ and to use Eq. (11) to estimate $\gamma_{k,t}$. Once the estimates of $\gamma_{k,t}$ are developed, the best predictors will be selected using the linear regression Eq. (8) and the stepwise approach.

Since the precipitation can fall in any of the three different stages, the estimation procedure includes three different cases. Let us define the following variables: $L_{i,0} = \min(r_{i,t}, t = 1, \ldots, n, i = 1, \ldots, 6)$, $U_{i,0} = \max(r_{i,t}, t = 1, \ldots, n, i = 1, \ldots, 6)$ and $q_i = (q_{i,1} + q_{i,2})/2$. It should be noted that the variable $r_{i,t}$ was introduced in Eq. (6).

(iv) Excess stage case

If for a given month the precipitation falls in the following range—$q_{i,2} \le r_{i,t} \le U_{i,0}$—then the estimates of $\gamma_{k,t}$ can be obtained as follows:

$$\hat{p}_{1,t} = 0.5 + \frac{0.5}{U_{i,0} - q_{i,2}}(r_{i,t} - q_{i,2}), \tag{19}$$

$$\hat{p}_{3,t} = (1 - \hat{p}_{1,t})\hat{p}_{1,t}, \quad \hat{p}_{2,t} = (1 - \hat{p}_{1,t})^2, \quad \text{and}$$

$$\hat{\gamma}_{k,t} = \ln\left(\frac{\hat{p}_{k,t}}{1 - \hat{p}_{k,t}}\right), \quad k = 1, 2, 3. \tag{20}$$

(v) Normal stage case

If for a given month the precipitation falls in the following range—$q_{i,1} \le r_{i,t} \le q_i$—then the estimates of $\gamma_{k,t}$ can be obtained as follows:

$$\hat{p}_{3,t} = 0.5 + \frac{0.5}{q_i - q_{i,1}}(r_{i,t} - q_{i,1}), \tag{21}$$

$$\hat{p}_{2,t} = (1 - \hat{p}_{3,t})\hat{p}_{3,t}, \quad \hat{p}_{1,t} = (1 - \hat{p}_{3,t})^2, \quad \text{and}$$

$$\hat{\gamma}_{k,t} = \ln\left(\frac{\hat{p}_{k,t}}{1 - \hat{p}_{k,t}}\right), \quad k = 1, 2, 3. \tag{22}$$

If for a given month the precipitation falls in the following range—$q_i \le r_{i,t} \le q_{i,2}$—then the estimates of $\gamma_{k,t}$ can be obtained as follows:

$$\hat{p}_{3,t} = 0.5 + \frac{0.5}{q_{i,2} - q_i}(q_{i,2} - r_{i,t}), \tag{23}$$

$$\hat{p}_{1,t} = (1 - \hat{p}_{3,t})\hat{p}_{3,t}, \quad \hat{p}_{2,t} = (1 - \hat{p}_{3,t})^2, \quad \text{and}$$

$$\hat{\gamma}_{k,t} = \ln\left(\frac{\hat{p}_{k,t}}{1 - \hat{p}_{k,t}}\right), \quad k = 1, 2, 3. \tag{24}$$

(vi) Scarce stage case

If for a given month the precipitation falls in the following range—$L_{i,0} \le r_{i,t} \le q_{i,1}$—then the estimates of $\gamma_{k,t}$ can be obtained as follows:

$$\hat{p}_{2,t} = 0.5 + \frac{0.5}{q_{i,1} - L_{i,0}}(q_{i,1} - r_{i,t}), \tag{25}$$

$$\hat{p}_{3,t} = (1 - \hat{p}_{2,t})\hat{p}_{2,t}, \quad \hat{p}_{1,t} = (1 - \hat{p}_{2,t})^2, \quad \text{and}$$

$$\hat{\gamma}_{k,t} = \ln\left(\frac{\hat{p}_{k,t}}{1 - \hat{p}_{k,t}}\right), \quad k = 1, 2, 3. \tag{26}$$

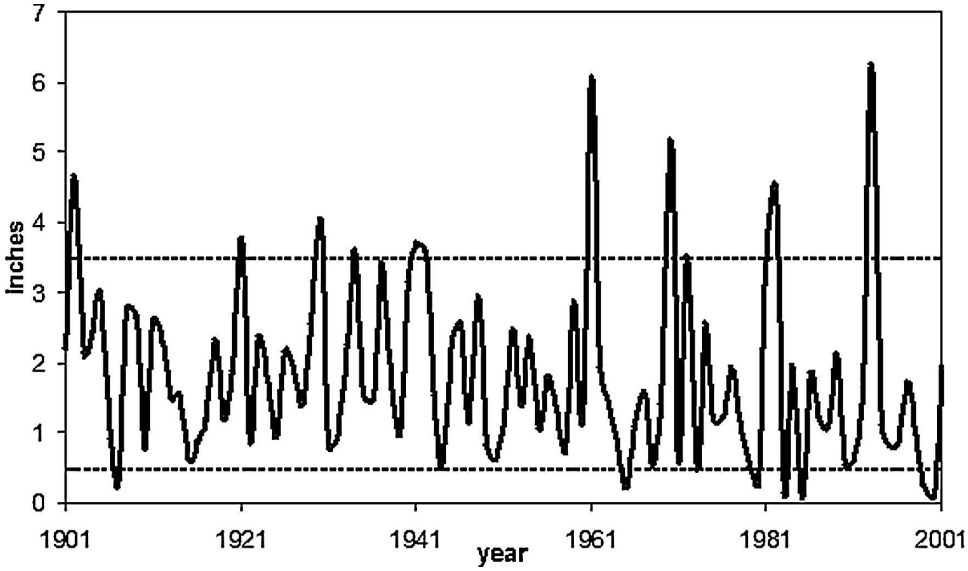

FIG. 3. Rainfall for the Mayagüez station to illustrate the numerical example. The lower horizontal line is the 10th percentile and corresponds to 0.52 in. of rain and the upper horizontal line is the 90% associated with 3.51 in. Any amount of rainfall that falls (during January at the Mayagüez station) between 0.52 and 3.51 in. represents a normal stage. Any observation that falls above 3.51 in. is declared an excess stage and any observation that falls below 0.52 in. is called a scarce stage.

2) NONLINEAR OPTIMIZATION ALGORITHM

The sequential quadratic programming (SQP) algorithm was used to estimate the parameters of the constrained nonlinear optimization problem given by Eqs. (14) and (15). The SQP algorithm approximates the objective function by a quadratic function and the constraints by a linear function. In this case, the underlying constraints are linear, that is, the approximation was computed only in the objective function. Once the initial point is selected and the approximation is complete, the quadratic problem is solved by using Wolfe's simplex method (Bazaraa et al. 1993; Winston 2004). The next point is considered a new initial point, and a new approximation is developed and solved. This process is repeated over and over in a sequential fashion until a maximum local is found. The implementation of the SQP algorithm is facilitated by using Matlab software (MathWorks 2000). Once the optimal solution is found, Eq. (7) is evaluated to predict the rainfall stage (excess, normal, or scarce).

4. Numerical example

A computer program was developed to model the rainfall processes of six stations located in PR. Ninety-one years of monthly observations (1901–91) were used as reference information and the last 10 yr of data (1992–2001) were employed to perform model validation. A numerical example is computed to illustrate the suggested methodology.

The proposed prediction scheme develops a model at every month and for every station. The Mayagüez station for January 1992 was arbitrarily selected to illustrate the model building and the optimization procedure. The algorithm starts by computing an initial point to solve the nonlinear problem. The initial point consists of finding an appropriate set of a values that guarantee convergence to a satisfactory local maximum. The 10 and 90%s were selected to define the stages of every month. Figure 3 shows 101 yr of precipitation at the Mayagüez station for the month of January; the upper horizontal line exhibits the 90% and the lower line shows the 10%. If the amount of rain is less than 0.52 inches per month, between 0.52 and 3.51 inches, or greater than 3.51 inches, the stage of the month will be considered scarce, normal, or in excess, respectively.

The procedure starts by computing the gamma values ($\gamma_{k,j}$) using Eqs. (19)–(26). Three linear models have been developed having the form expressed by Eq. (8). The VS algorithm identifies the predictors that best correlate with a given predictand. Since the available information is very limited, several lags and mathematical transformations are explored to increase the possibility of identifying appropriate models. Twenty-four lags and five mathematical transformations were explored (linear, squared, exponential, sinusoidal, and inverse). Thus, the total number of possible predictors is

TABLE 2. Initial and optimal point of the sequential quadratic programming algorithm. This table shows the optimization performance for the Mayagüez station for predicting January 1992 and shows the initial and optimal points obtained after conducting 184 iterations of the SQP algorithm. The first three columns show the parameters, initial point, and optimal point for the excess stage. Columns 4–6 and 7–9 show the same for the scarce and normal stages, respectively.

Excess stage			Scarce stage			Normal stage		
Parameter	Initial value	Optimal value	Parameter	Initial value	Optimal value	Parameter	Initial value	Optimal value
$a_{0,1,t}$	1.6598	−0.675 42	$a_{0,2,t}$	−2.3253	−2.0637	$a_{0,3,t}$	0.973 95	1.5498
$a_{1,1,t}$	−0.054 89	0.02397	$a_{1,2,t}$	0.434 45	0.124 04	$a_{1,3,t}$	−0.226 99	−0.021 99
$a_{2,1,t}$	−1.1242	−0.491 56	$a_{2,2,t}$	0.384 34	−0.005 07	$a_{2,3,t}$	−1.5695	−0.466 97
$a_{3,1,t}$	0.355 04	0.061 31	$a_{3,2,t}$	−0.267 49	−0.026 59			
			$a_{4,2,t}$	0.301 56	0.0619 15			
			$a_{5,2,t}$	0.064 614	0.025 443			
			$a_{6,2,t}$	−0.316 99	0.086 686			

17 original variables, 24 lags, and 5 mathematical transformations to each original variable. This turns out to be 2040 potential predictors. The inputs to the VS algorithm are one predictand ($\gamma_{k,j}$) and 2040 predictors. The algorithm develops a model for each month and uses 91 yr of data for model building (i.e., each variable has 91 observations). However, 88 observations are used to build the model since 3 observations were dropped because of the lag calculation. The variable selection algorithm creates groups of size $gs \leq 88/3 = 29.33$. The group size was 29 and the total number of groups of that size was 70; there was a single group of size 10. The stepwise algorithm and the multicollinearity subroutine were applied to identify the best variables. Thus, for the excess stage, 3 predictors out of the 2040 were identified as the winner predictors that best express the gamma values associated with the excess stage. Similarly, the routine for the VS algorithm was applied with different predictands and it identified 6 and 2 winner predictors for scarce and normal stages, respectively. The VS algorithm identified the following three models for the Mayagüez station to predict the rainfall stage for January 1992:

- Model for excess stage—

$$\begin{aligned}\gamma_{1,t} = {} & a_{0,t,1} + a_{1,t,1}e^{z_{3,t-20}} \\ & + a_{2,t,1}[\sin(z_{10,t-16}) + \cos(z_{10,t-16})] \\ & + a_{3,t,1}[\sin(z_{7,t-2}) + \cos(z_{7,t-2})] + \phi_{1,t}.\end{aligned} \tag{27}$$

The subscript t represents the time at which the prediction is made, that is, observations are known for time equal to 1, 2, . . . , t. Thus, $t = 12 \times 91 = 1092$ represents December 1991 and indicates that this is the most recent information that can be used to predict January 1992. The variable $z_{3,t-20}$ has two subscripts; the first one represents the number of the variable that is listed in Table 1 and the second subscript represents the time after the lag is implemented. The a coefficients from Eq. (27) represent the first set of coefficients of the initial point, and the complete initial point is given in Table 2. The variable $\phi_{1,t}$is the random noise with mean zero and constant variance.

- Model for scarce stage—

$$\begin{aligned}\gamma_{2,t} = {} & a_{0,t,2} + a_{1,t,2}z_{14,t-15}^2 \\ & + a_{2,t,2}[\sin(z_{15,t-9}) + \cos(z_{15,t-9})] \\ & + a_{3,t,2}[\sin(z_{5,t-19}) + \cos(z_{5,t-19})] \\ & + a_{4,t,2}z_{17,t-24}^2 + a_{5,t,2}z_{16,t-2}^2 \\ & + a_{6,t,2}z_{17,t-1} + \phi_{2,t}.\end{aligned} \tag{28}$$

- Model for normal stage—

$$\begin{aligned}\gamma_{3,t} = {} & a_{0,t,3} + a_{1,t,3}[\sin(z_{12,t-13}) + \cos(z_{12,t-13})] \\ & + a_{2,t,3}z_{7,t-11}^2 + \phi_{3,t}.\end{aligned} \tag{29}$$

The coefficients for a from these three models correspond to the initial point of the problem described by Eqs. (14)–(16). This optimization problem is solved by using the SQP algorithm. The initial and optimal points are shown in Table 2. The SQP algorithm converged after 184 iterations and the behavior of the ML function during the constrained optimization is given in Fig. 4. The ML function starts with a reasonable initial point reaching a value of −67.5 and after 64 iterations increases to a maximum of −54.6; however, at this point, the set of constraints is unfeasible. The algorithm continues for 184 iterations when the optimum and feasible point is reached with an ML function equal to −57. It should be mentioned that constraints are stabilized after 100 iterations, indicating that all the constraints are satisfied and have converged to a stable and reasonable value.

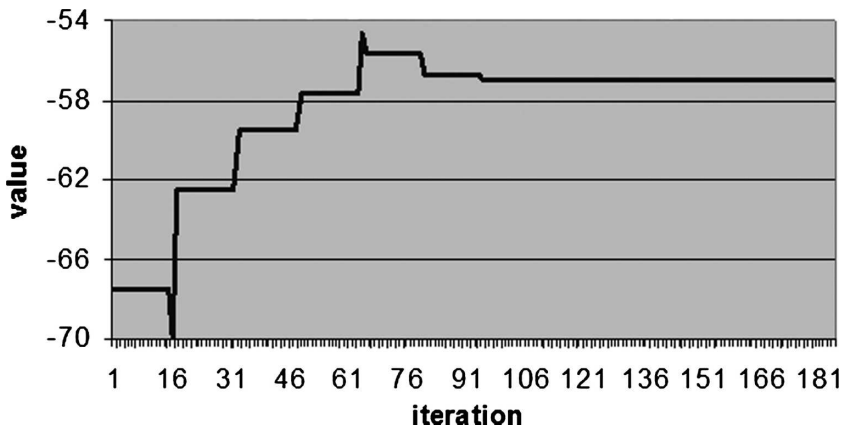

FIG. 4. The ML function. This figure represents the behavior of the objective function during the optimization process. The initial point starts at −68.5 and at the 64th iteration finds the maximum value. However, this solution is not feasible. The algorithm continues searching until it finally reaches the value of −57, and this is the feasible and optimal solution, which was obtained at iteration 184.

Once the optimal parameter set is obtained, the models [Eqs. (27)–(29)] are evaluated at time $t - g_j + L$, where t is the time when the forecast is made, g_j is the identified lag to the jth predictor, and L is the lead time. Table 3 shows the winner predictors evaluated at time $t - g_j + L$. Thus, Eqs. (27)–(29) are evaluated with optimal values of a at time $t - g_j + L$ and provide the following results: $\hat{\gamma}_{k,t+L} = \hat{\gamma}_{1,1093} = -1.1992$, $\hat{\gamma}_{2,1093} = -1.6835$, and $\hat{\gamma}_{3,1093} = 0.446\,16$. Finally, the predicted probability of being in a particular stage is computed by evaluating the following equations:

- Predicted probability of being in the excess stage in January 1992—

$$\hat{y}_{1,1093} = \frac{1}{1 + e^{-\hat{\gamma}_{1,1093}}} = \frac{1}{1 + e^{1.1992}} = 0.2316. \quad (30)$$

- Predicted probability of being in the scarce stage in January 1992—

$$\hat{y}_{2,1093} = \frac{1}{1 + e^{-\hat{\gamma}_{2,1093}}} = \frac{1}{1 + e^{1.6835}} = 0.1566. \quad (31)$$

- Predicted probability of being in the normal stage in January 1992—

$$\hat{y}_{3,1093} = \frac{1}{1 + e^{-\hat{\gamma}_{3,1093}}} = \frac{1}{1 + e^{-0.44616}} = 0.6097. \quad (32)$$

Based on the information from 1901 to December 1991, it is predicted that the Mayagüez station will be in the normal stage in January 1992, with a probability of 0.6097.

5. Cross validation

The prediction skills of the proposed scheme are assessed by means of cross validation. Essentially, 91 yr of data (1901–91) were used to develop forecast models and the last 10 yr (1992–2001) were used to validate the models. The validation of a model consists of comparing the forecasts with observations. Since the stage of a month is modeled by a multivariate Bernoulli distribution, the involved random variables are discrete and consequently the prediction accuracy of the stage can be studied by analyzing contingency tables. The validation process also includes the relative comparison of our model performances with a reference model that is known as the damped persistence model (Murphy 1992). The accuracy of the relative comparison is usually presented by using a specific skill score.

The typical scores that measure the accuracy of categorical forecasts are hit rate (H), probability of detection (POD), false-alarm rate (FAR), and bias (B). Hit rate is the fraction of the n forecasting occasions when the categorical forecast correctly anticipated the subsequent event or nonevent. Probability of detection is the likelihood that the event would be forecast, given that it occurred. The false-alarm rate is the proportion of forecast events that fails to materialize. Bias is the comparison of the average forecast with the average observations and is computed by the ratio of the average forecast with respect to the average of observations (Wilks 1995).

A three-way contingency table is used to validate the probability prediction scheme and includes the true stage and the predicted stage when the lead time pre-

TABLE 3. Predictors for probability stage. The original values of the predictors x are standardized by subtracting its mean and dividing by its standard deviation $z = (x - \bar{x})/s$. The z values are transformed and evaluated at time $t - g_j + L$, where $t = 1092$ and $L = 1$. The z values are multiplied by the optimal a values to estimate the gammas and to be able to estimate the probability of each stage. A routine to detect and fix the presence of outliers was implemented. This routine is critical, especially for the exponential and square transformation.

Excess stage		Scarce stage		Normal stage	
Predictor name	Predictor value	Predictor name	Predictor value	Predictor name	Predictor value
$e^{z_{3,t-20}}$	1.7329	$z^2_{14,t-15}$	0.3671	$\sin(z_{12,t-13}) + \cos(z_{12,t-13})$	−0.5075
$\sin(z_{10,t-16}) + \cos(z_{10,t-16})$	1.6320	$\sin(z_{15,t-9}) + \cos(z_{15,t-9})$	1.5337	$z^2_{7,t-11}$	2.3873
$\sin(z_{7,t-2}) + \cos(z_{7,t-2})$	3.8638	$\sin(z_{5,t-19}) + \cos(z_{5,t-19})$	3.6435		
		$z^2_{17,t-24}$	0.023		
		$z^2_{16,t-2}$	2.8439		
		$z_{17,t-1}$	4.2172		

diction is one month. When the true stage coincides with the predicted stage, the event is declared as a hit score; otherwise it is considered a failure. The POD and FAR scores include dichotomous yes/no forecast situations and consequently demand a conversion from three-way to two-way tables (Wilks 1995). Table 4 presents three-way contingency tables (one for each rainfall station) comparing the predicted stages versus the true stages. Transformation of the three-way tables into two-way contingency tables is not shown because of space limitations; however, results from the two-way tables such as POD, FAR, and B associated with each stage are also shown in Table 4. The outcomes from the six contingency tables were added to create a single contingency table that contains the total scores from the six stations. The 3×3 contingency table that contains the total scores will be called the total contingency table and reveals the overall performance of the forecasting scheme. Thus, the total contingency table provides the overall hit rate from the 6 stations during the 10-yr validation period. This hit rate value indicates that 85% of the time, the algorithm properly predicted the stages of the rainfall. The overall probabilities of detection were 0.52, 0.95, and 0.56 for the excess, normal, and scarce stages, respectively. The algorithm correctly predicts about 52% of the time when the stage is in excess, 56% of the time in scarce stage, and about 95% of the time when the process is in the normal stage. The false-alarm rates were 0.21, 0.12, and 0.28 for the excess, normal, and scarce stages, respectively. These results indicate that about 28% of the time or less, the algorithm predicts that the rainfall would be in a given stage when in reality a different stage has occurred. The bias scores were 0.65, 1.08, and 0.78 for the excess, normal, and scarce stages, respectively. Bias results indicate that the forecasts of the scarce and excess stages are underpredicted, whereas the forecasts of the normal stages show a small overprediction.

To compare the empirical probability model with a reference model, the forecasts were computed for the damped persistence model (Murphy 1992). The amount of predicted rainfall for each month was used to identify the prediction stage by using the 10th and 90th percentiles in the corresponding month. Based on the sequences of the predicted and observed stages, the 3×3 contingency tables for each station and the total contingency table from all stations were developed and analyzed.

Climatology forecasts were computed as the monthly average of observed rainfall. The period of time to calculate climatology changes from year to year since the average was computed from data starting in 1901 and ended up on the previous year of the prediction time, as indicated in Eq. (34). Persistence forecasts were computed by using either the observation of the previous month to be predicted or the observation made 12 months ago. The forecast for the damped persistence method uses a convex combination of the climatology and persistence forecasts. Since the observation of the previous month provides better forecasts than the one 12 months ago, the reported damped persistence method in this work is based on the previous month.

The damped persistence forecast was computed using the following expression (Murphy 1992):

$$\hat{r}_{\mathrm{dp},(\tau-1)12+m} = \alpha \hat{r}_{p,(\tau-1)12+m} + (1-\alpha)\hat{r}_{\mathrm{cli},(\tau-1)12+m}, \tag{33}$$

where

$$\hat{r}_{\mathrm{cli},(\tau-1)12+m} = \frac{1}{\tau-1}\sum_{i=1}^{\tau-1} r_{(i-1)12+m} \quad \text{and}$$
$$\hat{r}_{p,(\tau-1)12+m} = r_{(\tau-1)12+m-d} \tag{34}$$
$$\tau = 2, 3, \ldots, 101, \quad d = 1, 12,$$
$$m = 1, 2, \ldots, 12, \quad \text{and} \quad 0 \le \alpha \le 1,$$

TABLE 4. Three-way contingency tables describing the accuracy of the prediction scheme for each station and also a summary for all stations. Each three-way contingency table can be converted into 3 two-way tables, which were omitted. However, the results from these two-way tables are presented in terms of the following measurements: POD, FAR, and B, and these measurements are exhibited in the last four columns. The hit rate for each station is also presented in the last row of each station. The last table shows the three-way contingency table for the accumulation of scores from the six stations and the last four columns of the last table present results from the two-way contingency table describing the overall performance of the prediction scheme. The overall hit rate for the six stations is given in the last row of the table.

Station 1 Coloso		Obs				Excess	Normal	Scarce
		Excess	Normal	Scarce	POD	0.38	0.94	0.55
Model	Excess	3	3	1	FAR	0.57	0.14	0.14
	Normal	5	85	9	B	0.87	1.10	0.63
	Scarce	0	2	12			Hit rate = 0.83	
Station 2 Isabela		**Obs**				**Excess**	**Normal**	**Scarce**
		Excess	Normal	Scarce	POD	0.57	0.95	0.71
Model	Excess	8	0	0	FAR	0	0.07	0.54
	Normal	5	94	2	B	0.57	1.02	1.57
	Scarce	1	5	5			Hit rate = 0.89	
Station 3 Manati		**Obs**				**Excess**	**Normal**	**Scarce**
		Excess	Normal	Scarce	POD	0.45	0.95	0.64
Model	Excess	5	0	0	FAR	0	0.10	0.40
	Normal	5	90	5	B	0.45	1.05	1.07
	Scarce	1	5	9			Hit rate = 0.87	
Station 4 Manuanabo		**Obs**				**Excess**	**Normal**	**Scarce**
		Excess	Normal	Scarce	POD	0.58	0.95	0.55
Model	Excess	7	0	0	FAR	0	0.14	0.27
	Normal	5	84	9	B	0.58	1.11	0.75
	Scarce	0	4	11			Hit rate = 0.85	
Station 5 Mayagüez		**Obs**				**Excess**	**Normal**	**Scarce**
		Excess	Normal	Scarce	POD	0.40	0.95	0.54
Model	Excess	4	2	1	FAR	0.43	0.23	0.10
	Normal	6	71	15	B	0.70	1.23	0.60
	Scarce	0	2	19			Hit rate = 0.78	
Station 6 San Juan		**Obs**				**Excess**	**Normal**	**Scarce**
		Excess	Normal	Scarce	POD	0.80	0.95	0.45
Model	Excess	4	1	0	FAR	0.20	0.07	0.44
	Normal	1	99	6	B	1	1.02	0.82
	Scarce	0	4	5			Hit rate = 0.90	
All stations		**Obs**				**Excess**	**Normal**	**Scarce**
		Excess	Normal	Scarce	POD	0.52	0.95	0.56
Model	Excess	31	6	2	FAR	0.21	0.12	0.28
	Normal	27	523	46	B	0.65	1.08	0.78
	Scarce	2	22	61			Avg hit rate = 0.85	

where $\hat{r}_{\mathrm{dp},(\tau-1)12+m}$ is the amount of rainfall to be predicted by the damped persistence model at the τth year and in the mth month. Here, $\hat{r}_{\mathrm{cli},(\tau-1)12+m}$ is the amount of rainfall to be predicted by the climatology model at the τth year and in the mth month and r_t is the amount of rainfall observed in the tth month. Also, $\hat{r}_{p,(\tau-1)12+m}$ is the amount of rainfall to be predicted by the persistence model at the τth year and in the mth month. The subscript d represents either the observation from the previous month or from 12 months ago. It should be noted that the subscript for the station was dropped to simplify the notation. The convex coefficient (α) was obtained by minimizing the mean absolute prediction errors.

The sequence of stages identified by the climatology forecasts were compared with the observed sequence of stages during the validation period, and the associated contingency table is presented in Table 5. The predicted sequence of stages based on the damped persistence forecast was computed and also provided the same contingency table as the climatology forecasts. Table 5 shows that the overall hit rate for either the climatology or damped persistence forecast is equal to 0.77.

The relative comparison between the empirical probability and the reference models can be accomplished by using a skill score (SS) that relates the hit rates of the forecast methods. For instance, the hit rate between the empirical probability model and damped persistence model is given as follows:

$$\mathrm{SS}_{dp} = \frac{H_{\mathrm{emp}} - H_{\mathrm{dp}}}{H_{\mathrm{perf}} - H_{\mathrm{dp}}} = \frac{.85 - .77}{1 - .77} = 0.35, \qquad (35)$$

where H_{dp}, H_{perf}, and H_{emp} are the hit rates of the damped persistence, the perfect, and the empirical forecasts, respectively. This result shows that the empirical probability model exhibits an improvement of 35% over both the climatology and the damped persistence. The dynamic probability model has the capability of counting the number of times that a particular predictor belongs to the winner set. Thus, the contribution of each predictor is measured by counting the number of times that each variable belongs to the winner set. Results show that the selected meteorological indices can be used as proxy variables that may help predict the rainfall stages.

TABLE 5. Contingency table for climatology forecasts.

Station 1		Obs: Excess	Obs: Normal	Obs: Scarce	Hit rate
Forecast	Excess	0	0	0	0.75
	Normal	8	90	22	
	Scarce	0	0	0	
Station 2		**Excess**	**Normal**	**Scarce**	**Hit rate**
Forecast	Excess	0	0	0	0.83
	Normal	14	99	7	
	Scarce	0	0	0	
Station 3		**Excess**	**Normal**	**Scarce**	**Hit rate**
Forecast	Excess	0	0	0	0.79
	Normal	11	95	14	
	Scarce	0	0	0	
Station 4		**Excess**	**Normal**	**Scarce**	**Hit rate**
Forecast	Excess	0	0	0	0.73
	Normal	12	88	20	
	Scarce	0	0	0	
Station 5		**Excess**	**Normal**	**Scarce**	**Hit rate**
Forecast	Excess	0	0	0	0.63
	Normal	10	75	35	
	Scarce	0	0	0	
Station 6		**Excess**	**Normal**	**Scarce**	**Hit rate**
Forecast	Excess	0	0	0	0.87
	Normal	5	104	11	
	Scarce	0	0	0	
All stations		**Excess**	**Normal**	**Scarce**	**Hit rate**
Forecast	Excess	0	0	0	0.77
	Normal	60	551	109	
	Scarce	0	0	0	

6. Conclusions

One of the major contributions of the present research effort is to introduce probability models with empirical functions that predict rainfall level at a given station. These functions help to determine the parameters of the probability model with the use of historical meteorological events. Since meteorological events change over time and space, the probability model becomes dynamic. The multivariate Bernoulli distribution, introduced in this research, is used in conjunction with empirical models to relate the parameter vector of the probability model with observational atmospheric phenomena. The success of the probability model is highly dependent on the variable selection algorithm. If the appropriate variables are selected for an algorithm, the probability model will provide reasonable forecasts. An algorithm for variable selection is proposed in this paper where the amount of predictors is larger than the number of observations. The algorithm is based on three regression concepts: parsimonious principle, stepwise selection, and multicollinearity problem. The algorithm identifies the minimum number of variables that maximize the correlation between predictors and a predictand. The variable selection algorithm was used to identify the initial point to induce convergence in the sequential quadratic programming algorithm.

The model may not always predict the correct precipitation stage for a month. Ten years of cross validation, however, indicates that 85% of the time, the algorithm properly predicted the correct rainfall stage. The probability of detection was 0.52, 0.95, and 0.56 for ex-

cess, normal, and scarce stages, respectively. The false-alarm rate was 0.20, 0.12, and 0.28 for excess, normal, and scarce stages, respectively. Validation results indicated that the proposed methodology is a potential tool that may be used to predict the rainfall level in a given station. Results show that the selected meteorological indices can be used as proxy variables that can help predict rainfall stages.

Acknowledgments. This research has been supported by NASA EPSCoR Grant NCC5-595 and by the University of Puerto Rico Sponsored Research Office. The authors appreciate the technical support provided by the Climate Diagnostics Center and Drs. Robert S. Webb and Jon Eischeid who provided the 101 yr of monthly meteorological index data and rainfall data for the San Juan station. The authors acknowledge the collaboration of Israel Matos, meteorologist in charge of the National Weather Service in San Juan, PR. Helpful comments were provided by the reviewers, and one of them in particular gave a very thorough review along with useful recommendations.

REFERENCES

Bazaraa, M. S., H. D. Sherali, and C. M. Shetty, 1993: *Nonlinear Programming: Theory and Algorithms.* 2d ed. John Wiley and Sons, 638 pp.

Bickel, P. J., and K. A. Doksum, 1977: *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden-Day, 492 pp.

Box, G. E. P., and G. M. Jenkins, 1976: *Times Series Analysis: Forecasting and Control.* Holden-Day, 575 pp.

Brockwell, P. J., and R. A. Davis, 2002: *Introduction to Time Series and Forecasting.* 2d ed. Springer-Verlag, 434 pp.

Carter, M. M., 1999: A quantitative precipitation forecast experiment for Puerto Rico. Ph.D. dissertation, The Florida State University, 123 pp.

——, and J. B. Elsner, 1997: A statistical method for forecasting rainfall over Puerto Rico. *Wea. Forecasting,* **12,** 515–525.

Chiu, L. S., and B. Kedem, 1990: Estimating the exceedance probability of rain rate by logistic-regression. *J. Geophys. Res.,* **95** (D3), 2217–2227.

Comarazamy, D. E., 2001: Atmospheric modeling of the Caribbean region: Precipitation and wind analysis in Puerto Rico for April 1998. M.S. thesis, Department of Mechanical Engineering, University of Puerto Rico, Mayagüez, Puerto Rico, 112 pp.

Gahrs, G. E., S. Applequist, R. L. Pfeffer, and X.-F. Niu, 2003: Improved results for probabilistic quantitative precipitation forecasting. *Wea. Forecasting,* **18,** 879–890.

Gray, W. M., P. J. Klotzbach, and W. Thorson, cited 2005: Extended range forecast of Atlantic seasonal hurricane activity and U.S. landfall strike probability for 2005. [Available online at http://hurricane.atmos.colostate.edu/Forecasts/2005.]

Hurrell, J. W., 1995: Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science,* **269,** 676–679.

Janowiak, J. E., 1988: An investigation of interannual rainfall variability in Africa. *J. Climate,* **1,** 240–255.

Malmgren, B. A., A. Winter, and D. Chen, 1998: El Niño–Southern Oscillation and North Atlantic Oscillation control of climate in Puerto Rico. *J. Climate,* **11,** 2713–2717.

MathWorks, 2000: Optimization toolbox for use with Matlab: User's guide. MathWorks, Inc., 331 pp.

Montgomery, D. C., E. A. Peck, and G. G. Vining, 2001: *Introduction to Linear Regression Analysis.* 3d ed. John Wiley and Sons, 641 pp.

Mood, A. M., F. A. Graybill, and D. C. Boes, 1974: *Introduction to the Theory of Statistics.* 3d ed. McGraw-Hill, 564 pp.

Murphy, A. H., 1992: Climatology, persistence, and their linear combination as standards of reference in skill scores. *Wea. Forecasting,* **7,** 692–698.

Pandit, S. M., and S. M. Wu, 1983: *Time Series and System Analysis, with Applications.* John Wiley and Sons, 586 pp.

Reklaitis, G. V., A. Ravindran, and K. M. Ragsdell, 1983: *Engineering Optimization: Methods and Applications*. John Wiley and Sons, 684 pp.

Sohn, K. T., J. H. Lee, S. H. Lee, and Ch. S. Ryu, 2005: Statistical prediction of heavy rain in South Korea. *Adv. Atmos. Sci.,* **22,** 703–710.

Taylor, M. A., D. B. Enfield, and A. A. Chen, 2002: Influence of the tropical Atlantic versus the tropical Pacific on Caribbean rainfall. *J. Geophys. Res.,* **107,** 3127, doi:10.1029/2001JC001097.

Thompson, D. W. J., and J. M. Wallace, 1998: The Artic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophys. Res. Lett.,* **25,** 1297–1300.

Trenberth, K. E., 1984: Signal versus noise in the Southern Oscillation. *Mon. Wea. Rev.,* **112,** 326–332.

Ward, M. N., and C. K. Folland, 1991: Prediction of seasonal rainfall in the northeast of Brazil using eigenvectors of sea surface temperature. *Int. J. Climatol.,* **11,** 711–743.

Wei, W. W. S., 1990: *Time Series Analysis: Univariate and Multivariate Methods.* Addison-Wesley, 478 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

Wilson, F. W., 2004: Probabilistic forecasts of cloud impacts at San Francisco International Airport. *Proc. 20th Int. Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, 18.1.

Winston, W. L., 2004: *Operations Research: Applications and Algorithms*. 4th ed. Brooks/Cole, 1418 pp.