A Statistical Method for Forecasting Rainfall over Puerto Rico

M. M. CARTER AND J. B. ELSNER

Department of Meteorology, The Florida State University, Tallahassee, Florida

(Manuscript received 5 January 1996, in final form 14 April 1997)

ABSTRACT

Using results from a factor analysis regionalization of nontropical storm convective rainfall over the island of Puerto Rico, a statistical methodology is investigated for its potential to forecast rain events over limited areas. Island regionalization is performed on a 15-yr dataset, while the predictive model is derived from 3 yr of surface and rainfall data. The work is an initial attempt at improving objective guidance for operational rainfall forecasting in Puerto Rico. Surface data from two first-order stations are used as input to a partially adaptive classification tree to predict the occurrence of heavy rain. Results from a case study show that the methodology has skill above climatology—the leading contender in such cases. The algorithm also achieves skill over persistence. Comparisons of forecast skill with a linear discriminant analysis suggest that classification trees are an easier and more natural way to handle this kind of forecast problem. Synthesis of results confirms the notion that despite the very local nature of tropical convection, synoptic-scale disturbances are responsible for prepping the environment for rainfall. Generalizations of the findings and a discussion of a more realistic forecast setting in which to apply the technology for improving tropical rainfall forecasts are given.

1. Introduction and motivation

Tropical convection is notoriously difficult to forecast. The value of such forecasts is large, however, due to the potential for flooding and mudslides. This is particularly important in the Tropics where rainfall can be locally intense. A conditionally unstable atmosphere and an abundance of low-level heat and moisture combine with forcing mechanisms, such as sea-breeze fronts, to explain the contingency of tropical rainfall. As an example, one of the greatest problems facing weather forecasters in Hawaii is the prediction of heavy rainfall and its associated flash floods (Kodama et al. 1995).

Here we demonstrate a technology that holds promise in providing useful objective guidance for operational forecasters predicting tropical rainfall. The procedure involves a recently developed modification of the standard tree-structured classification method. Classification trees have been successfully applied to the problem of forecasting lake-effect snowfalls (Burrows 1991). The present test case is based on data from the eastern third of the island of Puerto Rico.

Currently there is little in the way of objective guidance to aid forecasters in the prediction of tropical convection over limited spatial scales. In Puerto Rico, for instance, a WSR-88D Doppler radar is used in an anecdotal capacity but is not continually in operation at the San Juan Weather Service Forecast Office (WSFO). Furthermore, high-resolution mesoscale numerical models are not available to forecasters in San Juan to provide guidance on a convective scale. The finest resolution model is the 29-km Eta Model, which covers the island with only eight grid points (S. Bennett 1996, personal communication). It is hoped that the model developed in this study will advance mesoscale forecasting. The utility of purely objective guidance is the independence from forecast experience or the skill of the forecaster (Ramage 1993). This can help stabilize the overall operational forecast performance of a WSFO in the event of staff turnover.

The paper is divided into two main parts. First a summary of the important results of a factor analysis regionalization of convective rainfall over Puerto Rico is given, followed by an example of how the technology of classification trees can be used to build a statistical prediction model. Specifically, a description of the available data is given in section two, followed in section three by details of the rainfall regionalization. Section four is a description of the predictors chosen for the forecast model. Section five contains details of the classification tree used with results of a limited prediction study presented in section six. The paper ends with a summary and a discussion in section seven.

We stress that this work is preliminary and it is meant to alert the operational forecast community to the utility of classification trees for developing objective forecast guidance. It is by no means the final word and, as is

Corresponding author address: Matthew M. Carter, Department of Meteorology, The Florida State University, Tallahassee, FL 32306-3034.
E-mail: carter@huey.met.fsu.edu

FIG. 1. Stations in Puerto Rico that record rainfall on an hourly basis.

usual in statistics, it is often best to try several different empirical approaches.

2. Data

The primary interest of this study is summertime convective rainfall. Therefore, only the months of May through September are considered. The study is based on two principal datasets: hourly rainfall from a network of stations and conventional hourly surface observations.

The U.S. National Climate Data Center records data for 22 stations in Puerto Rico on an hourly basis. Rainfall is collected in Fischer–Porter gauges distributed throughout the island, represented by the dots in Fig. 1. Because the gauges automatically record their contents, they may be placed in remote and mountainous regions of the island with little maintenance. All hourly rainfall amounts are in tenths of an inch, except for San Juan WSFO and Benavente-Hormigueros, which report in a resolution of one-hundredth of an inch. The rainfall data record includes the summer months for the years 1973–88, for a total of 55 080 h. Hours for which rainfall amounts were not reported are removed from the original dataset, as are hours influenced by hurricanes and tropical storms. Tropical cyclones produce widespread, torrential amounts of rain. In 1996, Hurricane Hortense produced catastrophic flooding in the small rivers and tributaries of Puerto Rico. Because the spatial and temporal scales of rainfall associated with tropical cyclones are resolved by several forecast models at the Tropical Prediction Center, it is felt that the skill of our algorithm for predicting such rainfall would not approach that of the dynamical model products. For this reason, hurricanes and tropical storms are removed from the dataset.

The forecast region for the predictive model developed in this study is the populous eastern third of the island. This designation is not entirely heuristic; it represents three convective regions of the island that exhibit similar diurnal rainfall frequencies, as will be shown in section 3. Although all rainfall recording stations are incorporated into the factor analysis that regionalizes the island, only the 11 stations that compose the eastern third of Puerto Rico are used for designing a prediction algorithm.

Hourly surface data for San Juan WSFO and Roosevelt Roads were retrieved from National Center for Atmospheric Research dataset ds472.0. The surface data encompass summer months May through September for the years 1977–79. Although the San Juan record is quite comprehensive, Roosevelt Roads surface data since 1980 does not contain mean sea level pressure, information used in building this prediction model.

Fifteen summers (55 080 h, May through September, 1973–88) of hourly rainfall data are used in regionalizing the island of Puerto Rico. The surface dataset is not this extensive, so the hourly rainfall dataset used to construct the prediction algorithm was pared to match that of the surface data (3 yr). Next, we choose one hour from each day [0800 AST (Atlantic standard time)] to initialize the statistical model. The potential maximum number of hours we can use to build our prediction model is 459 (153 summer days times 3 yr). Rainfall parameters, such as 12-h area-wide total, are calculated as possible predictors for day in the developmental sample. Then surface and rainfall data are matched chronologically so that for each day in the 3-yr period a complete data line is available. Any day that has missing information is discarded. All 12 h of rainfall data previous to 0800 AST must be present to fully complement the initialization hour. Each piece of missing persistence data effectively eliminates 1 of the 459 days.

Data was more likely to be missing at Roosevelt Roads than at San Juan, curtailing the number of initialization hours. Therefore, from a possible 459 initialization hours, 125 are used in the model building phase of our study.

Each month in the 3-yr period is represented by an initialization hour. Each of the 3 yr is well represented: 1977 has 30 initialization hours, 1978 has 53, and 1979 has 42. In part because we only have 125 prediction hours, we verify the stability of our results through cross validation, described in section 6.

3. Regionalization

The island of Puerto Rico is on a horizontal scale of a hundred kilometers. Synoptic-scale phenomena occur on a scale of a thousand kilometers, an order of magnitude larger. Large-scale midlatitude frontal passages do not occur during the summertime in Puerto Rico. Therefore, regions on the island that exhibit distinct variance signatures seek to capture rainfall forcings on the mesoscale. Such phenomena include, but are not limited to, sea breezes, mountain-top convection, orographic rain, and standing gravity waves. Passing easterly waves may also be included in this category. Though on the mesoscale, the horizontal scale of sum-

mertime rain phenomena is sufficiently large to encompass multiple stations. The average distance of proximal stations is only 10.7 km. Therefore, it is reasonable to suggest that several stations may share a common rainfall variance signature. This provides the impetus for identifying common regions of convective rainfall on this rather small tropical island.

We identify regions of mesoscale rainfall variance through the analysis of variance technique of factor analysis. We present the salient features of factor analysis while a complete description of the factor analysis regionalization is given in Carter and Elsner (1996). A more rigorous treatment of the factor analysis model is also presented in appendix A of this paper.

In contrast to the commonly employed principal component analysis, factor analysis starts with the assumption of an underlying basic model for the data. This model is given as

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\Psi},$$

where $\mathbf{\Sigma}$ is the population covariance matrix, $\mathbf{\Lambda}$ is the matrix of common factor loadings (the superscript T denotes the matrix transpose), and $\mathbf{\Psi}$ is the matrix of covariances of the specific factors. Shared variance among two or more rainfall stations is called communality and appears as a component in the matrix $\mathbf{\Lambda}$. For instance, Several coastal stations may exhibit rainfall variance due to a sea breeze. This communality is manifest in one component of $\mathbf{\Lambda}$. One of those coastal stations may further display a rainfall variance due to a very localized forcing, such as a nearby mountain peak. This peak does not affect the other stations, nor can it explain away the shared sea-breeze effect. It is an additive forcing of the rainfall variance for that particular station. This added variance appears in the vector ψ_i, which is the specific factor component of $\mathbf{\Psi}$ for station *i*. Each station has a unique component in $\mathbf{\Psi}$, while each common region has a component in $\mathbf{\Lambda}$. The goal is to determine the common factor loadings that relate individual stations to a shared variance signature. The loadings will determine to some extent to which region a particular rainfall station belongs based on covariance relationships between all other stations. This is accomplished by a spectral decomposition of the sample correlation matrix computed from the hourly rainfall amounts at each of the 22 stations.

A key decision to make in any factor analysis is how many common factors are necessary to best describe the covariance relationships among the variables. It is important that the sampling method for selecting the hourly rainfall is stable with respect to adjacent eigenvalues. A first-order approximation is given by perturbing the empirical orthogonal functions and calculating error bars for each estimated eigenvalue. If the error bars of adjacent eigenvalues overlap, then ''effective degeneracy'' occurs, with one estimation of the eigenvalues leading to a particular linear combination, and a second estimation leading to another (North et al. 1982). Overlapping will occur when the upper bound of an eigenvalue error interval exceeds the lower bound of its previous eigenvalue error interval. In other words, if $(\lambda_i + \delta\lambda_i) > (\lambda_{i-1} - \delta\lambda_{i-1})$, where λ_i and $\delta\lambda_i$ are the eigenvalue and eigenvalue shift, respectively, for station *i*, then effective degeneracy may occur. For our sample size of 55 080, only the error bars of eigenvalues 21 and 22 overlapped. Even if we restrict the degrees of freedom to 2295 by treating each day as independent, only eigenvalues 21 and 22 had an overlap in their error bars. Since it is not pragmatic to characterize the variance of 22 rainfall stations with 20 regions, we develop a more stringent criteria to choose the number of factors in our regionalization.

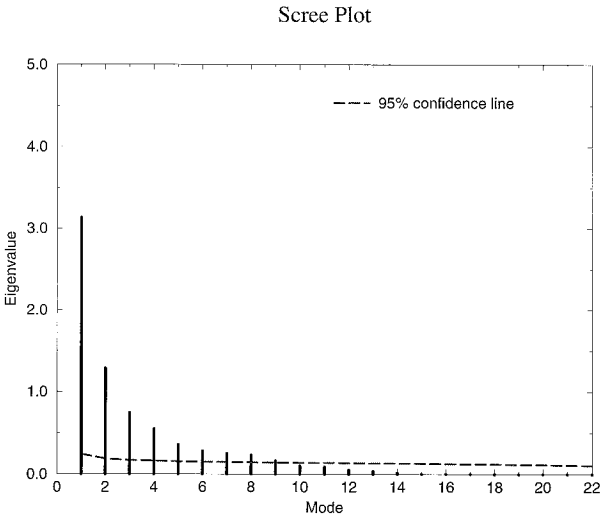

FIG. 2. Scree plot showing the 22 eigenvalues of the sample correlation matrix. The dashed line represents the 95% significance line from a Monte Carlo simulation of white noise. Nine of the 22 eigenvalues are above the significance line.

A Monte Carlo procedure that provides an upper bound on the number of statistically significant factors is used. The eigenvalues are plotted in Fig. 2. By spectrally decomposing 100 randomly generated surrogate rainfall data matrices, we choose the fifth largest eigenvalue for each mode to represent the 95% significance level (Overland and Presendorfer 1982; Elsner and Tsonis 1991). This significance level, which is based on the assumption of white noise, is shown as the dashed line in Fig. 2. The leading nine original data eigenvalues exceed this significance level and thus provide an upper bound on the number of factors to retain. Because we found nine modes to be significant with respect to white noise, each rainfall recording station does not by itself represent a unique rainfall region (if it did, all 22 modes would be significant). Stations may be grouped into regions, as long as the number of regions does not exceed nine. The significance test provides an important ''first guess'' as to how many factors should be considered in the analysis.

Additionally, we employ an orthogonal rotation that

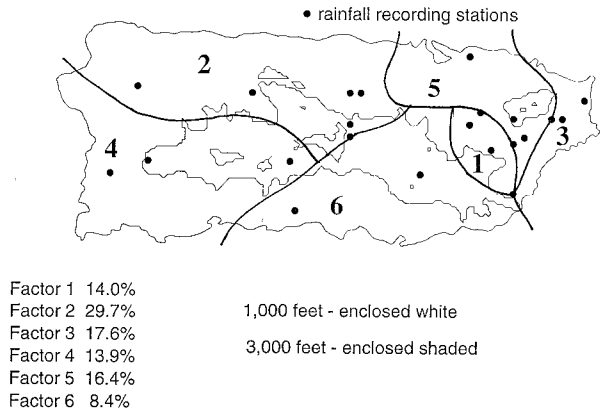

FIG. 3. Geographic regionalization of Puerto Rico based on a factor analysis (with an orthogonal rotation of the loadings) of summertime convective rainfall. Data are hourly rainfall amounts from 22 stations (dots) for the months of May–September over the years 1973–88 with hours influenced by tropical storms and hurricanes removed. Percentages reflect the relative contribution of each region to the hourly rainfall variance on Puerto Rico.

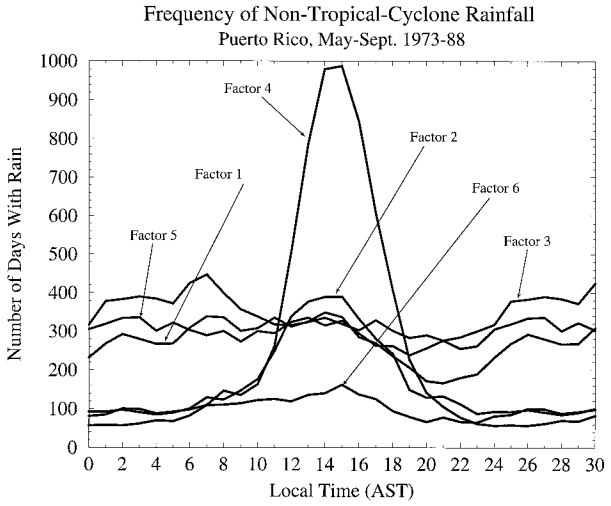

FIG. 4. The number of rainfall events for the summer months of May–September over the period 1973–88 is plotted for each hour of the day. Hours 24 through 30 represent ''wraparound'' times corresponding to 0000 through 0600 AST. A rainfall event is defined as any recorded amount of 0.25 cm or more occurring at any station within a region. Rainfall associated with tropical storms and hurricanes is excluded.

has the property of conserving the inner product of the loading vectors (columns of $\mathbf{\Lambda}$) and geometrically represents a rigid rotation about the coordinate axis (Kreyszig 1993). Rotation helps to reveal simple structure in the data. The rotated factor loadings may be plotted pairwise on an x–y axis (not shown). Simple structure occurs when the pairs align along the coordinate axes, indicating that a station is loading heavily on one factor and lightly on the rest (Carter and Elsner 1996). We choose a varimax rotation, performing the calculations through the FROTA subroutine of the International Mathematical and Statistical Library (IMSL).

Using the white noise significance test and orthogonal rotation as guides, we now proceed to regionalize Puerto Rico. We apply the factor analysis model for each m in the intervals 1 through 9 and carefully examine the factor loadings. We want to find an m for which all 22 stations optimally load on only one factor. It is the magnitude of the factor loadings that will determine the regions, since they are contributing to the common variance.

We call stations that load heavily on two or more common factors ''freeloading'' stations since they are free to load on more than one factor. Stations that do not significantly load on any common factor are called ''homeless'' stations, since they cannot be placed in a region based upon simple threshold criteria. We call the sum of freeloading and homeless stations the ''nonsingularity sum.'' If every station loaded above a predetermined threshold on only one factor, this sum would be zero and little subjectivity would be required. This is what we strive to achieve. In such a case, the only subjectivity lies in our initial choice of a loading threshold. More common, however, are analyses where some stations load on more than one common factor, while others load on none at all. In other words, the nonsingularity sum is almost always greater than zero. In our study, we found that a factor analysis with a choice of m = six factors provided the smallest nonsingularity sum. Based upon our loading threshold criteria, only three stations had to be placed subjectively because of their nonsingularity. For the three nonsingular stations, we draw the line on (or very close to) their locations. The regionalization using six common factors is shown in Fig. 3.

The six regions point to important physical mechanisms that force precipitation over the island and indicate that the factor analysis model is sensitive to variations in weather regimes (Carter and Elsner 1996).

Since our goal is to identify homogeneous rainfall regions to be used as forecast model targets, we seek characteristics of the rainfall regions that make the development of predictive algorithms tractable. We begin by examining the diurnal variability of precipitation in each of the six regions. This is done by considering the empirical probability of measurable precipitation for each hour. Figure 4 shows the frequency of rainfall (again excluding rainfall from tropical storms and hurricanes) for each hour of the day for each of the six regions.

Factors 1, 3, and 5, characterized by low-amplitude frequency maxima and minima, show small hourly variability. The ratio of maximum frequency to minimum frequency is greatest for region 5 and is no greater than 2.2:1. More importantly, the frequency maxima tend to occur in early morning. These factors correspond to regions on the island's eastern third. Factors 2, 4, and 6, on the other hand, have definitive afternoon frequency maxima. The ratio of maximum frequency to minima

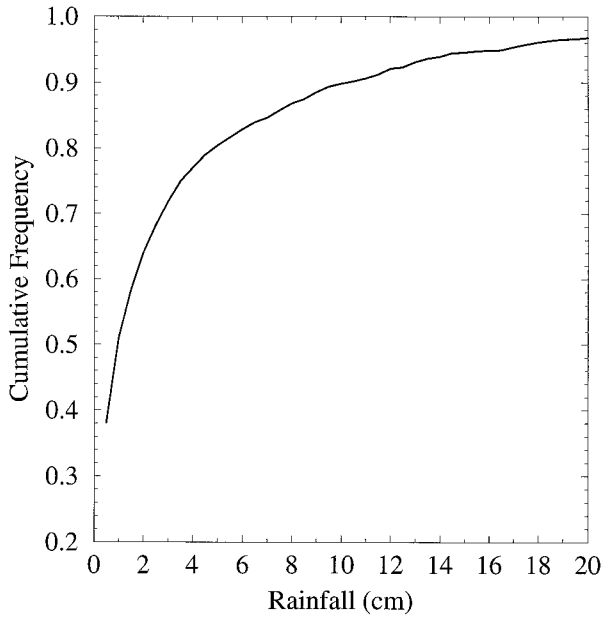

FIG. 5. Cumulative relative frequency distribution of 12-h total precipitation from the 11 rainfall stations in the eastern third of Puerto Rico. The period considered is from 0800 to 2000 AST during the months of May–September for the years 1973–88 excluding rainfall from tropical storms and hurricanes.

frequency exceeds 2.5:1 for all three of these regions and for region 4 is greater than 17:1. Factors 2, 4, and 6 compose the western two-thirds of the island. We distinguish this dichotomy first by the temporal occurrence of the frequency maxima and second by the amplitude of these maxima.

Based on their hourly frequency signatures, the island may be further separated into two larger regions: a western and an eastern region. Summarizing, the common factor analysis divided Puerto Rico into six distinct regions of hourly rainfall variability, and the empirical hourly rainfall probability reveals a dichotomy of diurnal variability. The forecast target for building the prediction model is the eastern ''superregion,'' comprised of convective regions 1, 3, and 5.

Since we target the 11 station total precipitation over the 12-h period from 0800 to 2000 AST for developing the prediction algorithms below, here we present a brief rainfall climatology of this region. Only days on which all stations were reporting each hour over the period 1973–88 (May–September) were used in the climatology. There are a total of 1529 rainfall values. The minimum 12-h total is zero and this occurred 24% of the time. The maximum total is 75.0 cm. The average 12-h rainfall total over the eastern superregion is 3.5 cm with a standard deviation of 6.9 cm. The cumulative frequency distribution is shown in Fig. 5.

4. Selecting the predictors

We now attempt to build a prediction algorithm for daytime convective rainfall. On many conditionally unstable days, convective rainfall is on a small enough scale that it may not reach a rain gauge in the vicinity of the shower. Only rainfall that falls into the gauge will appear in the record, even if it is raining heavily nearby. This aspect of convective rainfall makes it very difficult to predict. This is especially true in Puerto Rico during the summer because variations in such variables as temperature, dewpoint, and wind direction are small on a diurnal basis. Predicting for an entire rainfall region instead of a single point is more reflective of the prevailing convective forcing.

TABLE 1. Predictor variables used to build the statistical forecast models in this study.

Label	Predictor variable
X_1	San Juan surface wind (u component) m s^{-1}
X_2	Roosevelt Roads surface wind (u component) m s^{-1}
X_3	San Juan sea level pressure anomaly mb
X_4	Roosevelt Roads sea level pressure anomaly mb
X_5	San Juan sea level pressure anomaly tendency mb 12 h^{-1}
X_6	Roosevelt Roads sea level pressure anomaly tendency mb 12 h^{-1}
X_7	Eleven station composite: % stations reporting rain during previous hour
X_8	Eleven station composite: past 12-h total rainfall

Large-scale disturbances are often responsible for creating an environment favorable for rain in the Tropics. Predictors based on a physical understanding are the most natural candidates for producing a successful forecast model (Ramage 1993). We choose a set of eight variables that we feel are important in setting up a favorable convective environment. This is done by first considering the following potential predictors.

- *Surface winds*: Tropical waves and diurnal sea breezes change the wind speed and direction, and may have a significant effect on convective rainfall in Puerto Rico (Gere Gallup, personal communication).
- *Sea level pressure anomalies*: Tropical waves and their attendant moisture are often characterized by an inverted trough in the pressure field. Mean sea level pressure is calculated for each hour of the day throughout the entire data record. The appropriate mean is subtracted from each hourly sea level pressure value to give the anomaly. In this way, the semidiurnal pressure oscillation is removed.
- *Sea level pressure anomaly tendencies*: The pressure tendency determines whether the tropical wave is approaching or departing the region. The net change in sea level pressure anomaly is calculated over the 12 h previous to 0800 AST.
- *Past rainfall*: Persistence can often be an important parameter in short-range weather forecasting.

These variables are considered as potential predictors. From them, and based on data availability from both San Juan and Roosevelt Roads, we extract eight predictors (Table 1) for building the prediction model. We note that this is a small subset of variables and includes

no upper-air, no satellite, and no radar information. It does, however, provide a starting point for evaluating prediction technologies. We add that a stepwise linear regression (not shown) using data from San Juan and a larger set of potential predictors also identified the above variables as the most important predictors. The exclusion of upper-air humidity variables in the predictor set implies that their predictive information is contained within the persistence rainfall variables.

Since we seek to predict daytime convective rain, our predictor data is taken only at 1200 UTC (0800 AST). This is the time at which we initialize the model. The predictand data is taken from rainfall over the period from 1200 UTC (0800 AST) to 0000 UTC (2000 AST).

5. Classification trees

To develop an effective set of prediction rules for forecasting convective rainfall, we desire that the method have several characteristics to ensure its usefulness and validity. Among the most important of these considerations is that the methodology allow for statistical significance testing by way of cross validation, allow for nonfunctional relationships between predictor variables and the predictand, and provide useful and easily interpretable results. Methods such as linear programming do not allow for statistical validation of the results, while purely statistical methods like regression and discriminant analysis do not easily allow for nonfunctional relationships.

Therefore, to create a set of prediction rules for convective rainfall, we experiment with a statistical classification algorithm known as partially adaptive classification trees, or PACT (Shih 1993). PACT unifies the multivariate statistical methodology of linear discriminant analysis (LDA; Mardia et al. 1979) and tree-structured classification methods (CART; Breiman et al. 1984). As will be discussed, PACT combines the advantages of both methodologies and meets the desired criteria specified above. We note that among classification methods the algorithm chosen here is not unique; however, it is quite simple to implement and yields satisfactory results for the purposes here. Readers wishing to investigate other classification methods are encouraged to refer to Breiman et al. (1984) and Hand (1981).

Here, Q is the universe of J disjoint subsets, $A_1, \ldots, A_j$, that may be expressed by $Q = \cup_j A_j$. A classifier is a portion of Q into these subsets such that for all $x \in A_j$ the predicted class is j. A classifier can be constructed based on past experience. For example, suppose that heavy afternoon showers are common when morning pressure tendencies are substantially positive. In this case Q is the universe of all pressure tendencies with A_2 being the pressure tendencies less than some value x_c and A_1 being pressure tendencies greater or equal to x_c. Then for all $x \in A_1$, the predicted class is 1 or heavy rain. In general x will be multidimensional, so it will be a vector $\mathbf{x}$, and for our problem $\mathbf{x}$ will be of fixed length so that the data have standard structure.

A brief review of LDA and analysis of variance (ANOVA; Casella and Berger 1990) is needed in order to understand how the PACT algorithm creates its classification rules. Linear discriminant analysis is a multivariate statistical technique that seeks to classify an observation into a group or category according to the observed values of several associated predictor variables. The choice of a linear discriminant function (LDF) depends upon the nature of the data involved. The most commonly used LDF assigns group classifications by using a generalized distance function (the Mahalanobis distance) that measures the distance of the values of the predictor variables corresponding to an observation to the means of those predictor variables for each classification group (Mardia et al. 1979). An observation is then assigned to that group for which its distance measure to the group mean (the centroid) is the smallest.

ANOVA is a technique to determine how much a measured predictand varies according to different group classifications and to ascertain the corresponding statistical significance (Casella and Berger 1990). This methodology uses least squares techniques to estimate the sources of variances so that a single test statistic (the F statistic) can indicate the statistical significance of the variance caused by the group classifications. PACT also uses Levine's test, which measures how much variance is caused by the group classifications. Levine's test is quite robust and formally tests for equality of group variances in continuously valued data. Like ANOVA, Levine's test also creates the F statistic as its single test statistic.

PACT itself functions by emulating the decision trees created by CART. A decision tree is a set of sequential rules that one follows in order to classify an observation; the name itself comes from the appearance of the rules as written on a sheet of paper, which is somewhat similar in appearance to a flow chart. Within a decision tree, each time a decision (or classification rule) is to be performed, we are at what is called a decision node. The result of the decision, true or false, shunts the decision into a choice of two other nodes, which themselves may be either another decision node or what is known as a terminal node. A group classification is assigned for each terminal node. To make a prediction, we begin at the first decision node (the top of the tree) and ultimately finish in a terminal node at some part of the tree.

As an example of this, and to illustrate the major advantage of PACT over LDA, refer to Fig. 6. Here, an artificially created dataset shows a separation of light and heavy rainfall events by wind speed and sea level pressure anomaly tendencies.

An optimal set of classification rules would stratify the variable space in the simplest manner possible so that we could accurately predict every single observa-

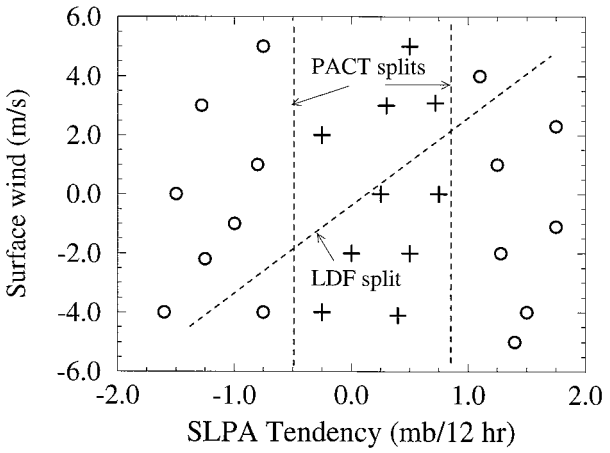

FIG. 6. An artificial example showing the difference in results between the PACT and LDA classification schemes. Pluses indicate the occurrence of heavy rain and circles indicate the occurrence of light rain. The splits from the PACT algorithm are represented by two vertical lines and the split from the LDA algorithm is represented by the diagonal LDF line.

tion. For the case here, the stratification by rainfall is nonfunctional; that is, there is no linear discriminant function that can divide this region into the proper subregions for light versus heavy rainfall.[1] In other words, since we have two categories (light and heavy rain) and two predictor variables, LDA is limited to separating the regions by the best straight line that can be drawn in the plane.

In contrast, PACT is not limited by the nonfunctional relationship. The PACT algorithm here produces a decision tree with two decision nodes and three terminal nodes and achieves a 100% accurate classification. At the first decision node the question is asked, ''Does the observation have a surface pressure anomaly tendency of less than -0.5 mb 12 h^{-1}?'' If yes, than we are shunted to a terminal node with a light rain label. Otherwise, we are shunted to the second decision node that determines whether or not pressure tendency is greater than 0.8 mb 12 h^{-1}. If yes, then the observation is shunted to a terminal node and classified as light rain, otherwise it is shunted to the other terminal node and classified as heavy rain. Note that the PACT algorithm ignores surface wind entirely in its decision process as it contains no useful information, while the LDA spuriously uses wind in its region separation, separating the regions with a line that is constructed as a linear combination of the two predictor variables. If wind had contained useful information, then PACT would have also used this at a decision node, but in a univariate fashion. That is, PACT separates the regions in a univariate fashion so that we do not have to evaluate linear combinations. This is particularly useful in high-dimensional datasets. Of course, the exmple provided here is for illustrative purposes only, and the results should not be interpreted to mean that PACT is universally superior to LDA. It is not.

Here we use PACT software developed by Y.-S. Shih at the University of Wisconsin–Madison. We make no attempt to compare results with other tree-based classification algorithms, such as CART. Details of the method are provided in appendix B.

6. Results

a. PACT model development

We perform an experiment that resembles the situation a forecaster might face as a way to compare the tree-based classification model with the more familiar discriminant analysis. Our purpose is twofold: to describe the information available from the classification tree forecast model and to demonstrate that classification trees can be a more powerful and more natural way to develop prediction strategies for this type of forecast problem. We do not, however, advocate that they will always perform better than other methods, and it is typically prudent to try other methods as well.

The experiment involves the forecast of categorical precipitation amount in the 12-h interval defined above (0800 to 2000 AST). The forecasts are issued based on data up through 0800 AST so the forecast has a zero-hour lead time. We note that this is all done on historical data so that the forecasts are actually hindcasts. The forecast target is the 11-station total precipitation over the 12-h period divided into two categories of light (less than 2 cm) and heavy (greater or equal to 2 cm). The choice of cutoff between light and heavy is arbitrary but is motivated by the desire for nearly equal prior probabilities.

As previously mentioned, because data for the eight predictors were not always available, the prediction experiment consisted of only 125 cases. With each case there were values of all eight predictors and a 12-h rainfall total with no missing hour at any of the 11 stations. The relatively small number of cases makes it necessary to use cross validation to assess the model accuracy (see Michaelson 1987; Elsner and Schmertmann 1994). Since the 125 cases are effectively chosen at random throughout the 3-yr period, we can treat each case as independent and use a hold-one-out cross-validation strategy (see Elsner and Schmertmann 1994).

Figure 7 shows a typical regression tree from the cross-validation prediction experiment. The circles represent decision nodes and the boxes terminal nodes. Adjacent to each decision node is the variable chosen by the algorithm for a binary split (refer to Table 1 for the variable label and description). For example, the first decision is whether or not the previous 12-h rainfall (X_8) is less than 1 cm. If the answer is yes, then follow the

[1] Strictly speaking, a relation $y = f(x)$ is functional if every x produces a unique y. If not, the relation is nonfunctional.

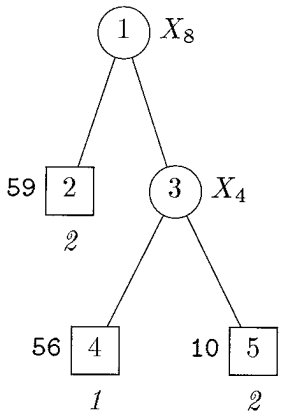

FIG. 7. Decision tree diagram for predicting categorical precipitation over Puerto Rico. Decision nodes are circles, and terminal nodes are boxes marked below the box as 1 for heavy rain and 2 for light rain. Nodes are labeled sequentially inside. The used in the decision split is indicated to the right of the decision node, and the of cases falling in each decision group is located to the left of the terminal node.

left branch to the terminal node marked two. Terminal node two indicates that a prediction of light rain should be made for the next 12 h. Alternatively, if the previous rainfall exceeds 1 cm, then the right branch is chosen leading to another decision node. This decision node involves the sea level pressure anomaly at Roosevelt Roads (X_4). If the pressure anomaly is less than 1.6 mb, then follow left to the terminal node marked four. This node indicates that a prediction of heavy rain should be made. If the pressure anomaly exceeds 1.6 mb, then follow to the right terminal node marked five and a prediction of light rain is made.

The decision rules chosen by PACT for splitting the rainfall events can generally be found to make physical sense. For example, the decision to forecast light rain when the previous 12-h rainfall was very light is known as persistence. Also, the classification tree procedure produces a ranking of the relative importance of the predictors used in the tree construction. Table 2 lists the rankings for the variables used to make the tree in Fig. 7.

Not surprisingly, the variable picked as most important was the past 12-h rainfall over the region. Other important variables included sea level pressure anomalies and anomaly tendencies. The surface wind component was found to be the least important. Although all variables are ranked, only the most significant ones are used in the model. As such, a high-ranking variable may not appear in the final model if it is not statistically significant. Note that although X_3 outranks X_4 slightly, it is supplanted by X_4 in the decision tree *after* X_8 is chosen as the first decision node. That is, some of the information that ranked X_3 above X_4 is contained in X_8.

As expected, it appears that the predictable component of convective rainfall is the one associated with tropical wave activity. Westward-moving tropical waves with their pressure anomalies are a common feature of the summertime weather regime of Puerto Rico. These waves are often difficult to detect from satellite imagery owing to their lack of convection as they make their way across the Atlantic. However, as they reach Puerto Rico, and in response to daytime surface heating of the island landmass, they can initiate widespread convection that may last for a few days. Thus both persistence and surface pressure anomalies are useful predictors of heavy rainfall.

TABLE 2. Rank of the relative importance of the prediction variables used in constructing the classification tree. Higher numbers imply greater importance.

Predictor variable	Ranking
X_1	40
X_2	41
X_3	85
X_4	84
X_5	81
X_6	64
X_7	81
X_8	100

b. Model comparison and cross validation

We compare the performance of the PACT algorithm with that of a linear discriminant analysis using a hold-one-out cross-validation strategy and the Heidke skill score (HSS) as a measure of forecast performance. Discriminant analysis is closely related to regression since the object is to calculate a linear function that best separates the groups (light vs heavy rain) on the basis of a number of predictors measured for all of the individuals in each group.

Furthermore, we compare PACT against a persistence forecast. If the hours previous to 0800 AST were raining lightly, then light rain becomes the persistence forecast. If the overnight hours experienced heavy rain, then heavy rain is the persistence forecast.

The PACT, discriminant analysis, and persistence models are used to forecast the occurrence or not of heavy rainfall ($\geq$2 cm) over the eastern region for the 12 h ending at 2000 AST from data up through 0800 AST using the eight previously described predictors from San Juan and Roosevelt Roads. Table 3 shows the results with N the number of cases, E the number of

TABLE 3. Model comparisons using HSS and the approximate correlation coefficient based on a two-category categorical forecast. Here, N is the number of cases, E is the number of cases correct if climatology is used, and H is the number of cases correct if the forecast model is used. The approximate correlation coefficient is based on Barnston (1992).

Model	N	E	H	HSS	Approx. correlation
Discriminant analysis	125	71	90	0.352	0.55
Persistence	125	71	86	0.278	0.43
PACT	125	71	94	0.426	0.62

correct if simply always forecasting light rain (climatology), and H the number of correct using the forecast algorithm. The HSS, computed as

$$\text{HSS} = \frac{H - E}{N - E},$$

is 0.352 for the linear discriminant analysis, 0.278 for persistence, and 0.426 for the PACT. Using the approximate relationship between HSS and correlation coefficient for a two-category decision (Barnston 1992), we find that the PACT algorithm provides a respectable correlation between the actual and forecast events of 0.62.

It should be emphasized that the results are based on the prior selection of a 2-cm cutoff and have not been cross validated with respect to this criterion. As such, there is an in-sample bias in the skill of the PACT model that does not exist for the discriminant model. Caution should therefore be exercised in assigning physical significance to the analysis results.

Again, we hasten to add that this result is based on a single experiment and may not accurately represent the limitations (or power) of the PACT methodology for these kinds of forecast problems. Nevertheless, we feel that because the tree-based methods allow for nonfunctional relationships between predictors and the predictand, they hold promise for developing useful statistical forecast guidance for such events. Further, since the classification trees do not depend on any strong distributional assumptions, they can be used on data with highly nonsymmetric distributions, like convective rainfall.

7. Summary and discussion

The purpose of this study was to demonstrate the potential of developing useful guidance products from a modern statistical technology for forecasting tropical convection. A factor analysis was consulted as a way to rationally divide the island of Puerto Rico into coherent convective rainfall regions. An area encompassing the eastern third of the island, where the diurnal rainfall signal is weak, was chosen as a target area for an attempt at forecasting heavy rain events. Total rainfall from 11 stations within this region over a 12-h period (0800–2000 AST) was used as the predictand. Instead of predicting for exact rainfall amounts, the challenge was to predict the category of light (<2 cm) or heavy rainfall.

A partially adaptive classification tree (PACT) algorithm that unifies tree-structured classification and discriminant analysis is used to predict categorical precipitation with some success. The Heidke skill score (HSS) exceeds 0.4 and compares favorably with a score of 0.35 using linear discriminant analysis and 0.28 using persistence. Comparisons in forecast model skill were done using a hold-one-out cross validation. The important predictors, as indicated by the classification tree, point to the importance of the large-scale environment in forcing the very local scale convective storms. Results should be treated with some care since they are based on a single experiment with only a limited number (125) of cases.

The approach outlined in this paper should be more effective in combination with other models. For instance, the prediction of forecast errors from a dynamical forecast model or a model output statistics forecast system should be investigated. Moreover, at this point we have made no attempt to determine the value of the tree-based classification model within the framework of a cost–loss ratio situation (see e.g., Murphy 1977; Murphy and Ehrendorfer 1987). Of course the most successful tropical forecast models will make use of all available observations including satellite and radar imagery.

Acknowledgments. We thank Shawn Bennett, scientific operations officer at the San Juan WSFO, for his assitance in providing current model guidance information and geographical insight on the island of Puerto Rico. We also acknowledge G. S. Lehmiller for his help with the PACT algorithm and some of the statistical interpretations. Support for this work came from NOAA through the Cooperative Institute on Tropical Meteorology. We also acknowledge the National Center for Atmospheric Research Data Support Service. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or its subagencies.

APPENDIX A

The Factor Analysis Model

The common factor analysis model begins with the underlying model

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Psi},$$

where $\boldsymbol{\Sigma}$ is the $p \times p$ population covariance matrix, $\boldsymbol{\Lambda}$ is the $p \times m$ matrix of common factors, $\boldsymbol{\Lambda}^{\mathrm{T}}$ is its $m \times p$ transpose, and $\boldsymbol{\Psi}$ is the $p \times p$ matrix of covariance of specific factors ϵ_i. There are p rainfall stations and m factors. The relationship between the common factors and specific factors is given by

$$\mathbf{X} - \mu = \boldsymbol{\Lambda}\mathbf{F} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ is the hourly rainfall vector, $\mathbf{F}$ is the vector of common factors, and μ is the mean hourly rainfall for each station, where the expected values of $\mathbf{F}$, $\boldsymbol{\epsilon}$, and $\mathbf{X} - \mu$ are zero. Because we begin with a common factor model, we assume that the specific factors are independent of each other. The population covariance matrix may be normalized and estimated by the sample correlation matrix $\mathbf{R}$. Thus, we avoid one variable with large variance unduly influencing the determination of factor loadings (Johnson and Wichern 1982). Individual

elements in **R** between stations i and k are normalized by their individual variances,

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}},$$

whose covariance of hourly rainfall amounts between stations i and k at hour j is given by

$$s_{ik} = \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k),$$

where $\bar{x}_i$ and $\bar{x}_k$ are sample hourly rainfall means. When the matrix of specific factors Ψ is subtracted from the correlation matrix R, we are left with a dispersion matrix, which may be decomposed into its eigenvalues Γ and its eigenvectors Δ. The relationship between the common factor loadings, the eigenvalues, and eigenvectors is given by

$$\mathbf{\Lambda} = \mathbf{\Gamma}\mathbf{\Delta}^{-1/2}.$$

The initial estimate of factor loadings is noniterative and assumes the specific factors are equal to $(1 - m/2p)(1 - r_{ik}^2)$, where m is the number of factors in the model. These estimates are refined using an unweighted least squares iterative technique. All calculation estimates are performed by the subroutine FACTR in the IMSL.

For further information about the convective rainfall regionalization of Puerto Rico by factor analysis, please refer to Carter and Elsner (1996).

APPENDIX B

Details of PACT

PACT creates it classification rules by using a hybrid of several statistical methods. The procedure begins by creating an initial decision node and then adding further nodes as constrained by the tree growth parameters. Since it is possible to always create a 100% classification accuracy by completely partitioning the predictor space, a criterion is needed to determine the optimal tree size. Here, this was achieved by using a direct stopping rule and then maximizing the cross-validated classification accuracy as a function of the stopping rule. Since at this step the forecast algorithm is adjusted based on verifying observations, the procedure is called ''partially adaptive.'' A direct stopping rule stops the tree growth process once the number of observations remaining within a terminal node falls below a certain percentage threshold of the total number of observations. In other words, suppose a direct stopping rule of 6% was chosen. Then if the number of observations in a particular node is less than 6% of the initial total, the growth process is stopped for that node and it is assigned as a terminal node.

The algorithm functions are as follows. First, if the initial or any subsequent node has a sufficient number of observations, the algorithm performs an ANOVA on each potential predictor variable and selects the variable that has the most significant F statistic. To avoid ignoring variables that have a large degree of nonfunctional group separation, a Levine's test is also conducted to identify which variable has the largest inequality of variances caused by group classification. The F statistics for this test are also obtained. PACT selects the splitting variable for the decision node based on the variable having the largest F value over both test procedures.

Next, the algorithm performs a one-dimensional linear discriminant analysis, using the variable selected above. The decision rule for the decision node in question is created from the LDF, which partitions this node into two new (sub-)nodes. Finally, each of these nodes is checked to see if it has a sufficient number of observations, and the process repeats until all of the remaining nodes become terminal nodes, thus completing the tree.

The classification tree once completed allows for rather straightforward predictions. While the rules strictly create a binary decision, probabilities of assignments may also be estimated by one of two ways. One method is simply to note the observed classification error for the corresponding terminal node and calculate the group assignment probability as one minus the node misclassification error. Since this method ignores the actual values of the predictor variables outside of the classification cutoffs, another method consists of obtaining the corresponding group classification probabilities for the LDF used at each involved decision node and then using conditional probabilities to estimate the group assignment probabilities. Note that since LDA is technically a Bayesian classifier (Mardia et al. 1979), Bayesian prior probabilities may be utilized in the LDFs for each decision node. The PACT algorithm allows for this; however, we have not made use of prior probabilities in this study.

REFERENCES

Barnston, A. G., 1992: Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Wea. Forecasting,* **7,** 699–709.

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

Burrows, W. R., 1991: Objective guidance for 0–24-hour and 24–48-hour mesoscale forecasts of lake-effect snow using CART. *Wea. Forecasting,* **6,** 357–378.

——, M. Benjamin, S. Beauchamp, E. R. Lord, D. McCollor, and B. Thomsom, 1995: CART decision tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. *J. Appl. Meteor.*, **34,** 1848–1862.

Carter, M. M., and J. B. Elsner, 1996: Convective rainfall regions of Puerto Rico. *Int. J. Climatol.*, **16,** 1033–1043.

Casella, G., and R. L. Berger, 1990: *Statistical Inference.* Wadsworth, 650 pp.

Elsner, J. B., and A. A. Tsonis, 1991: A note on the spatial structure of the covariability of observed Northern Hemisphere surface air temperatures. *Pure Appl. Geophys.,* **137,** 133–146.

——, and C. P. Schmertmann, 1994: Assessing forecast skill through cross validation. *Wea. Forecasting,* **9,** 619–624.
Hand, D. J., 1981: *Discrimination and Classification.* John Wiley and Sons, 218 pp.
Johnson, R. A., and D. W. Wichern, 1982: *Applied Multivariate Statistical Analysis.* Prentice-Hall, 409 pp.
Kodama, K. R., G. M. Barnes, and T. A. Schroeder, 1995: Heavy rain forecasting for the southeast flank of the island of Hawaii. Preprints, *21st Conf. on Hurricanes and Tropical Meteorology,* Miami, FL, Amer. Meteor. Soc., 315–317.
Kreyszig, E., 1993: *Advanced Engineering Mathematics.* John Wiley and Sons, 1271 pp.
Mardia, K. V., J. T. Kent, and J. M. Bibby, 1979: *Multivariate Analysis.* Harcourt Brace and Company, 521 pp.
Michaelson, J., 1987: Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.,* **26,** 1589–1600.
Murphy, A. H., 1977: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.,* **105,** 803–816.
——, and M. Ehrendorfer, 1987: On the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Wea. Forecasting,* **2,** 243–251.
North, G. R., T. L. Bell, and R. F. Cahalan, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.,* **110,** 699–706.
Overland, J. E., and R. W. Presendorfer, 1982: A significance test for principal components applied to a cyclone climatology. *Mon. Wea. Rev.,* **110,** 1–4.
Ramage, C. S., 1993: Forecasting in meteorology. *Bull. Amer. Meteor. Soc.,* **74,** 1863–1871.
Shih, Y.-S., 1993: Tree-structured classification. Ph.D. thesis, University of Wisconsin—Madison, 161 pp. [Available from Theses Basement North, Memorial Library, University of Wisconsin—Madison, Madison, WI 53706.]