

# **Application of Neural Network-Oriented Pattern Classifiers to Estimate Missing Daily Rainfalls in the Subtropical Region**

Tae-Woong Kim and Hosung Ahn \*

South Florida Ecosystem Office - ENP/NPS

Homestead, FL 33030

File: Wrr\_Missing\_v4.doc (09/10/2004)

Be Submitted to *Water Resources Research*

---

\* Corresponding author, South Florida Ecosystem Office - ENP/NPS, 950 N. Krome Ave.,  
Homestead, FL, 33030, Telephone: 305-224-4215, Fax: 305-224-4147, Email:  
Hosung\_Ahn@nps.gov

Key words: Rainfall, Missing data, Pattern classification, Discriminant analysis, Neural networks

## **Abstract**

We developed a spatial daily rainfall model that is designed specifically to fill in missing rainfalls in the subtropical area where convective storms are dominant. Because the spatial pattern of daily rainfall occurrence is complex and nonlinear compared with those of frontal storms, we adopted neural network-oriented pattern classifiers to determine daily wet or dry (rainy or no-rainy) conditions. The model uses a two-step approach as most Markov-type daily rainfall models do: First, the classifiers differentiate daily wet or dry areas based on available recorded rainfalls, from which wet or dry condition at each missing site is determined. Then, a regression model estimates the amount of missing rainfall at each wet site determined at the first step. The following four neural network-based classifiers were tested using the measured rainfall data in South Florida: Levenberg-Marquardt Backpropagation, Automated Regulation Backpropagation, Learning Vector Quantification, and Probabilistic Neural Network approaches. The result reveals that the classifier based on a probabilistic neural network approach is superior to the others. Also, a stepwise regression performs better for estimating the amount of rainfalls than the other competing approaches we tested. We validated that the proposed model produces accurate and unbiased daily rainfall estimates.

## **1. Introduction**

Missing data are very common in most hydrologic records. They are due to bad weather condition, equipment mal-function, data contamination, or data processing errors. The incomplete data often lead to inconsistent and biased estimations. For instance, the sample covariance matrix computed from incomplete data is likely to produce negative eigenvalues that in turn leads to an erratic result in statistical analyses (Schneider, 2001). One could use only

complete data, but this could induce significant bias if missing data are not identically distributed. Also, many hydrologic models need seamless records to simulate continuous historical events. Thus, the first step in hydrologic analysis and modeling is to fill in missing gaps.

Various statistical methods have been developed to fill in missing values: First, the simplest method is to replace missing values with a mean, median, or mode of recorded data. This approach is often used in water quality data analyses. However, care should be taken to avoid insertion of bias, especially when data are not “Missing At Random” (Batista and Monard, 2003). Second, single or multiple linear regression models are most common in hydrology, where missing data are estimated based on the neighboring recorded data. This method often leads to underestimation of variance in space and time, if observations of two or more nearby sites are used (Salas, 1993). Third, time series models, such as autoregressive models, can be used to fill in missing observations if data are serially correlated or no other concurrent information from nearby sites are available (Salas, 1993). Forth, Batista and Monard (2003) examined the use of the k-nearest neighbor (KN) algorithm to make up for missing data. They demonstrated, from their experimental analyses using four Machine Learning Datasets from UCI Repository (<http://www.ics.uci.edu/~mlern/MLRepository.html>), that estimates based on the KN algorithm outperforms the simple mean or mode replacement method. The KN algorithm uses a mean of k-nearest neighbors as a best estimate. This method needs no explicit model to be built. However, this method is expensive because searching for the set of similar instances over the entire model domain is non-trivial. The estimated values by the KN algorithm induce unessential bias when the strong correlations exist among data.

As an alternative but more elaborate approach than the above methods, one could apply the expectation-maximization (EM) algorithm (Dempster et al., 1977; Schneider, 2001). This

method is a very powerful statistical technique designed specifically for incomplete data. The EM algorithm basically minimizes the error estimates of the likelihood function of a representative model for given data in an iterative manner. In order to estimate missing values and to estimate mean and covariance matrix, the EM algorithm iterates both expectation and maximization steps until the stopping criteria are satisfied. Usually, the EM algorithm underdetermines the conditional expectation of missing values given the available (complete) data. This is the case when data are severely missing so that the EM algorithm could not estimate adequate model parameters (Schneider, 2001). When the number of variables exceeds the number of records, the regularization method is used to make up the deficiency of the EM algorithm. Schneider (2001) tested the regularized EM algorithm with surface temperature data. He showed that the regularized EM algorithm leads to more accurate estimates of the missing values compared with conventional methods. However, daily rainfall data are an intermittent process that is not easily handled by the EM algorithm.

The alternating process of combination of rainy and no-rainy (zero value) sequences makes many statistical models impractical. For example, zeros in k-nearest neighbors cause a severe bias problem in estimating the neighboring values. They also result in a singular problem during matrix operations in an explicit model. The Markov-type models have been commonly applied to the intermittent daily rainfalls (Roldán and Woolhiser, 1982; Woolhiser and Roldán, 1982; Rajagopalan et al., 1996; Woolhiser, 1992; Wilks, 1998). The Markov-type rainfall models comprise of two components; the first part models the occurrence of rainfall based on the state of previous day – a typical first-order Markov process. The next part handles the distribution of rainfall amounts on rainy days independently from the first component. To fill in missing data, one could take advantage of concurrent neighboring records by modifying the conventional

Markov-type models. In the subtropical area like South Florida, convective storms are dominant. The corresponding pattern of daily rainfall occurrence in this region is complex and nonlinear compared with those of frontal storms. In fact, the preliminary analysis of rain data in the region indicates that the spatial dependency of rainfall occurrence is quite stronger than the serial dependence of rainfall occurrence.

Thus, the objective here is to develop a practical model to fill in missing daily rainfalls in South Florida. The model applies a pattern classification technique to determine the occurrence of rainfalls in space. The pattern classification used in the proposed model is expected to reduce the complexity of problems in estimating missing daily rainfalls. The next section describes briefly the rainfall monitoring network in South Florida and their statistical properties that are relevant to the structure of the proposed model.

## **2. South Florida Rainfalls**

The study area is located in and around of the Everglades National Park (ENP) in South Florida. The area is classified as the humid subtropical climate zone, having a hot, humid, and wet summer. The wet season extends from June through October, while the dry season is from November to next May. An average annual rainfall in the ENP is about 1400 millimeters, among which about 70 per cent are concentrated on the wet season. Summer rainfalls are in the form of convective local thunderstorms. Tropical storms and hurricanes also play a significant role in increasing wet season rainfalls (Winsberg, 2004).

Many local and federal governments including the ENP have been involving in the restoration of the Everglades under the name of the Comprehensive Everglades Restoration Plan (CERP) and other projects. The projects have been relying on various hydrologic models to

evaluate structural and operational alternatives. The models are often simulated at daily or less time steps because the surface water and groundwater systems in the area are directly connected and groundwater flows fast through a highly transmissible surficial aquifer (Tarboton et al., 1999; Wasantha Lal, 2001). These models use daily rainfalls as input. However, daily rainfalls recorded in the area have many short breaks that needed to be filled in on a scientific basis before using them in the models.

Currently, there are over 60 active rain gauges in the study area (Figure 1). Each station has different period of records. Most sites have data during the past 30 years, while couples of them have data back to early 1900s. Figure 2 shows the missing rate (the rate of missing days to recorded days) on each site. An average of missing rate is about 11 per cent, while a maximum rate is 32 per cent at western Florida Bay (station #60 in Figure 1). The missing rates during the wet and dry seasons are nearly identical. An average of rainy days per year is about 128 days (35%).

### **3. Simple Rainfall Occurrence Processes**

In general, a daily rainfall model is consisted of two independent components - rainfall occurrence and rainfall depth. The conventional models rely on a stochastic process such as a first-order Markov model to handle the rainfall occurrence. Let us describe the concept of a conventional Markov model and a linear classification model to help understand the proposed neural network models.

#### **3.1. Conventional Random (RND) Model**

The RND approach relies on a single site Markov process without considering

neighboring measurements. Rainfall occurrence is determined by the state transition probabilities – the probability of being a wet day from a previous wet day or from a previous dry day. Wet/dry condition ( $c_k(t)$ ) of  $k$ -th station at day  $t$  is determined based on a critical transition probabilities ( $p_c$ ) and generated random number ( $0 \leq u \leq 1$ ) from a uniform distribution (Roldán and Woolhiser, 1982; Hanson et al., 1994; Wilks, 1998).

$$c_k(t) = \begin{cases} 1(\text{wet}), & u_k(t) \leq p_c(t) \\ 0(\text{dry}), & \text{otherwise} \end{cases} \quad (1)$$

$$\text{where } p_c(t) = \begin{cases} Pr(c_k(t)=1/c_k(t-1)=0), & \text{if } c_k(t-1)=0 \\ Pr(c_k(t)=1/c_k(t-1)=1), & \text{if } c_k(t-1)=1 \end{cases}.$$

The transition probabilities may vary over the year. The non-homogeneous property can be considered in the model by allowing the transition probabilities to vary systematically over the year using a Fourier series reconstruction (Roldán and Woolhiser, 1982). The RND method is for a single site model rather than a spatial model as will be introduced here.

### 3.2. Classification-Based Models

Figure 3 depicts the structure of the proposed model. Each gauged rainfall value is dichotomized into wet ( $c_k(t)=1$ ) or dry ( $c_k(t)=0$ ) state depending on whether the rainfall is greater or less than a rainfall measurement error ( $=0.25$  millimeters), respectively. For each missing rainfall at a site on a day, a classifier determines whether the station is wet or dry. This step relies on the pattern classification of the spatial distribution of wet and dry stations, regardless of their magnitudes. The next step estimates the amount of rainfall if the station is turned out to be wet. The rainfall depth is estimated based on a multiple regression model using the concurrent neighboring measurements as independent variables. A preliminary data analysis shows that the amounts of rainfalls are remarkably correlated with the values at nearby stations

in most cases. Missing data in a given day are distributed randomly. Thus, the regression model to estimate rainfall depth needs to be constructed and estimated for each missing station separately, which is non-trivial. We adopted a preliminary classification process that spots and eliminates dry stations in advance in order to reduce such a laborious job. This process is turned out to be quite effective.

The purpose of the classification is to partition the spatial domain into two rainfall occurrence states - rainy or no-rainy. A pattern classification identifies the spatial boundaries of two states, from which the state for each missing station is classified. There are many statistical classifiers that evaluate the values of objects based on their spatial statistical properties. The next subsection will first describe a simple linear discrimination analysis to help understand the concept of statistical discrimination. Then, four neural network-based nonlinear classifiers applicable to our rainfall models will be introduced.

### 3.3. Linear Discriminant Analysis

Let us define that a discriminant function,  $g(\mathbf{x})$ , generalizes the pattern classification, where  $\mathbf{x}$  is a feature vector that is the spatial information for rainfall occurrence in this case. Rainfall occurrence is a binary process such that a classifier assigns a feature vector to class  $c_1$  (wet) when if  $g(\mathbf{x}) > 0$ , or  $c_2$  (dry) when  $g(\mathbf{x}) \leq 0$ . Linear discriminant analysis (LDA) is based on a combination of simple linear discriminant functions as

$$g(\mathbf{x}) = \mathbf{w}'\mathbf{x} \tag{2}$$

where  $\mathbf{w}$  is a vector of weights that maximizes the ratio of between-groups variance to within-groups variance (Cooley and Lohnes, 1971; Duda et al., 2000; Dudoit et al., 2000). Since  $g(\mathbf{x})$  is linear, the decision surface is a hyperplane which divides the space into two sub-spaces. The

linear discriminant function,  $g(\mathbf{x})$ , is given as an algebraic measure of the distance from  $\mathbf{x}$  to the hyperplane (Duda et al., 2000).

The LDA approach is simple to understand and easy to compute. However, the classification result of this method is too simple to characterize the complex wet-dry boundaries created by convective local thunderstorms in South Florida. For example, there are 24 stations available for pattern classification of rainfall occurrence on June 9, 1990 (Figure 4). In this case, we need a complex classifier using higher order nonlinear functions such as the curved line on Figure 4 in order to draw an appropriate wet-dry boundary.

One may adopt a very sophisticated classifier that may lead to a perfect classification for a particular training set. However, this could often lead to overfitting and results in poor performance on new or future patterns. Technically, this phenomenon is called the vulnerability to variability of feature objects (Duda et al., 2000). With an appropriate nonlinear function and a proper generalization of pattern classification, an optimal decision boundary can be established to have sufficient accuracy of performance of both training and future data sets. Taking into account for this consideration, it deems that multilayer neural networks could provide an optimal solution to the classification of the occurrence of South Florida rainfalls.

#### **4. Structure of the Proposed Model**

The proposed model relies on a neural network classifier to model the condition of rainfall occurrence at missing stations. That is, the rainfall occurrence at missing site is determined based on the patterns of occurrence of multiple measurements from nearby stations. We propose to use the multilayer neural network (MNN) approaches (Figure 5). The MNN is a massive parallel-distributed processing system. This system consists of many artificial neurons

that are highly interconnected by weights (Haykin, 1994; Duda et al., 2000). Owing to its innate nonlinear property and flexibility for modeling and training (ASCE Task Committee, 2000), MNNs have been widely applied for solving the problems of function estimation, time series analysis, pattern recognition, and system control.

In pattern classification problems, two-layer networks without hidden layer may be enough to form the decision boundaries of a binary problem. The decision boundaries in this case are composed of hyperplanes that have similar limitation to the LDA. Therefore, various MNNs, which have hidden layers as shown in Figure 5, have been developed to construct nonlinear decision boundaries for the pattern classification problems. The decision boundaries are achieved by the explicit discriminant function given as:

$$g_k(\mathbf{x}) = \hat{c}_k = f_o \left( \sum_{j=1}^m w_{kj} \cdot f_h \left( \sum_{i=1}^n w_{ji} x_i + w_{jo} \right) + w_{ko} \right) \quad (3)$$

where  $m$  and  $n$  are the numbers of neurons in the hidden and the input layers, respectively.  $f_h$  and  $f_o$  are activation functions in the respective hidden and the output layers,  $w_{ji}$  and  $w_{kj}$  denote the input-hidden layer weight and the output-hidden layer weight, respectively, and  $w_{jo}$  and  $w_{ko}$  represent biases for the hidden and the output layers, respectively. We propose to use one of the following four neural network classifiers that are described briefly on the following subsections. More details on these approaches are found in Demuth and Beale (2000) and Duda et al. (2000).

Once the state of rainfall occurrence is determined, the amount of rainfall is estimated day by day. A prior examination of our rainfall data reveals that the spatial correlation is more significant than the temporal one. Thus we prefer to use a multiple regression approach rather than a conventional Markov process. The spatial regression type model is given by:

$$\hat{y} = f(\mathbf{x}) + \varepsilon \quad (4)$$

where  $\hat{y}$  is the dependent variable representing the depth of rainfall at a missing station,  $\mathbf{x}$  a predictor matrix consisting of recorded rainfall depths at nearby stations, and  $\varepsilon$  is a prediction error. We propose either stepwise regression, principal component regression, or MNN based approaches to model the rainfall depth.

#### 4.1. Levenberg-Marquardt Backpropagation (LMB)

MNNs are trained to approximate any functions with optimizing the network weights,  $\mathbf{w}$ . To train MNNs, the backpropagation algorithm has been applied successfully to solve difficult and diverse problems. Standard backpropagation is a gradient descent algorithm. The network weights are moved along the negative of the gradient of the performance function,  $\mathbf{e} = \mathbf{c} - \hat{\mathbf{c}}$ . In this study,  $\mathbf{c}$  and  $\hat{\mathbf{c}}$  are vectors of measured and estimated class values, respectively. Then, the iterative training process is given as;

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \mathbf{H}^{-1} \mathbf{g} \quad (5)$$

where  $\mathbf{w}_k$  is the weight matrix at iteration  $k$ ,  $\mathbf{g}$  is the gradient ( $\mathbf{g} = \frac{\partial \mathbf{e}}{\partial \mathbf{w}}$ ), and  $\mathbf{H}$  is the Hessian

matrix ( $\mathbf{H} = \frac{\partial^2 \mathbf{e}}{\partial \mathbf{w}^2}$ ).

The standard backpropagation is often too slow to converge in practice. One of faster algorithms developed so far is the LMB algorithm. The LMB was designed to accelerate the training by eliminating the need to compute the Hessian matrix in Eq. 5. When the mean squared error is used as a performance function ( $\mathbf{e}$ ), the Hessian matrix can be approximated by the Jacobian matrix ( $\mathbf{J}$ ), as  $\mathbf{H} = \mathbf{J}^T \mathbf{J}$ . The corresponding gradient can be computed as  $\mathbf{g} = \mathbf{J}^T \mathbf{e}$ . The LMB appears to be fast and efficient for training moderate-sized feedforward neural networks.

## **4.2. Automated Regulation Backpropagation (ARB)**

Overfitting could be of concern in MNN trainings. The network is trained with a set of data to have very small value of error. When new data set are presented, however, the network occasionally yields large errors. This is because the network has a remarkable memory on a training data set but not for general or new situations. The estimated weights in this case are tuned only for the specific training set. The overfitting occurs when insufficient training data are given or when the variability of data is significant. The overfitting problem may happen in daily rainfall modeling by various reasons. In the northwest corner of the area, the network density is relative low. Even in the dense gauging area, the number of the training data set is insufficient when there are many missing data. For instance, there are only 24 stations available on June 9, 1990 (Figure 4). These undesirable circumstances make the estimated weights in the network sensitive to the network configuration data and feature values.

One of ways to prevent the overfitting is to make regularization of a network. That is, the performance function is modified by adding a regularization parameter which combines the weighted mean square errors instead of using the mean squares of weights as in the previous case. The ARB algorithm determines a set of optimal regularization parameters during training the network until the effective weights and biases converge.

## **4.3. Learning Vector Quantization (LVQ)**

This algorithm trains competitive layers automatically by clustering input vectors. The clusters that are found in competitive layers are dependent only on the distance between input and weight vectors. The LVQ network learns to classify input vectors into target classes chosen

by user in a supervised manner. The LVQ network has two layers - competitive layer (hidden layer) and linear layer (output layer). The competitive layer learns to cluster input vectors by assigning “1” to the winner neurons, or “0” to the others. The winner neurons are associated with the most positive element of the net input of a competitive layer. The elements of net input are the negative of the distances between input vector and the weight vector. The competitive layer then forms a cluster in which the similar input data are gathered. The linear layer transforms the competitive layers into target classifications (user input) through a training process that adjusts the weights in a competitive layer.

#### **4.4. Probabilistic Neural Networks (PNN)**

This approach consists of two layers. The first layer (hidden layer) links between input neurons and respective pattern neurons that are in turn connected to class neurons in the second layer (output layer). In the first layer, the distance vector from the input neurons to the pattern neurons is computed. The second layer sums these contributions for each class of inputs to produce its net output of a vector of probabilities. A compete transfer function is used in the second layer to pick the maximum of these probabilities and assign “1” for that class. “0” is assigned for the other classes.

The PNN is trained by modifying the weights connecting the input neurons to the pattern neurons in a following way. The normalized pattern neurons are placed on the normalized input data such that  $w_{ji} = x_{ji}$ , where  $w_{ji}$  is the weight connecting the  $i$ th input neuron to the  $j$ th pattern neuron, and  $x_{ji}$  is the  $i$ th normalized input to the  $j$ th pattern neuron. If  $x_i$  is assigned to a  $c_k$  class, then,  $w_{kj} = 1$ , or otherwise  $w_{kj} = 0$ . After training, the network is fully connected between the input neurons and the pattern neurons. The pattern neurons are identically connected to the

class neurons. The trained network is then used for pattern classification. The PNN is straightforward and does not need an iterative training. The PNN always guarantees convergence when a set of sufficient training data is given.

## **5. Evaluation of Rainfall Occurrence Process**

This study tested four neural network-based classifiers (LMB, ARB, LVQ, and PNN), which are available as options on the MathWorks software (Demuth and Beale, 2000), for the classification step in the proposed model. Also tried are two simple approaches, namely RND and LDA as benchmarks. To test the classifiers in the proposed rainfall model, this study picked 10 arbitrary stations that have low missing rate (see Figure 1 and 2). Table 1 lists summary statistics as well as missing rate at each selective station. Furthermore, this study selected two spatial rainfall sets (both wet and dry seasons) of 100 randomly selected days for which both rainy and no-rainy sites are adequately mixed to form the spatial pattern of rainfall occurrence on each day. The days having both all rainy stations and all no-rainy stations were excluded on this selection. Then, gaps on each data set were created artificially and randomly to mimic the real data having missing records (a missing rate of about 10 percent), so that the estimated rainfall values by the proposed model are compared with the corresponding measured values.

### **5.1. A Number of Hidden Neurons**

In a pattern classification problem, the numbers of input neurons ( $n$ ) and output neurons ( $k$ ) are dependent on the dimensions of input and output. Hidden neurons link between input and output neurons. The number of hidden neurons ( $m$ ), which determines the accuracy of decision boundary, is a key factor to govern the complex of a network. However, there is no established

method to determine  $m$ , nor any straight relation between  $m$  and the dimension of classification (Duda et al., 2000). It is well known that a convenient rule of thumb is to choose  $m$  value such that the total number of weights in a network is roughly a tenth of the number of data set (Duda et al., 2000). Experimental studies revealed that the networks having fewer hidden neurons than input neurons have worked well (Maier and Dandy, 2000). The number of inputs of the network for the classification (described in the next section) is not sufficient for the above considerations. A large  $m$  produces complex decision boundaries and takes long time to train the network. On the other hand, a small  $m$  creates too rough decision boundaries to get an accurate pattern classification. The  $m$  should be chosen to make tread-off between the accuracy of classification and the generalization of the network. This study set  $m$  as a half of the number of training sets, that is,  $m$  is a function of the number of available stations in this case, to gain substantial accuracy and flexibility to the problem.

## 5.2. Input Features

In the proposed model, a classifier determines daily rainfall occurrence based on the feature values ( $\mathbf{x} = (x_1, \dots, x_{nf})$ ,  $nf$  is the number of features) that includes spatial covariates. The covariates to the rainfall could be geographical information (x-y coordinates) or physical features such as concurrent water level data. In South Florida, simulated daily water levels (surface ponding depth) from 1965 to 2000 on the 2 mile by 2 mile square grids are available (SFWMD, 1999). The water levels are simulated with spatial rainfalls that are interpolated into grid values using the Triangular Irregular Network (TIN) method. The co-located model-drive water levels are highly correlated to daily rainfall and thus be an excellent input feature to model the rainfall occurrence. Three combinations of input features were examined here: Option 1 employs x- and

y-coordinates of the station as input features (for example, at station #1,  $\mathbf{x}_1 = (x_1, y_1)$ ,  $nf=2$ );

Option 2 adopts the Option 1 features plus model-driven ponding depths ( $z$ ) corresponding to the station ( $\mathbf{x}_1 = (x_1, y_1, z_1)$ ,  $nf=3$ ); and Option 3 only uses a model-driven ponding depth ( $\mathbf{x}_1 = (z_1)$ ,  $nf=1$ ).

### 5.3. Comparison of Alternative Approaches

The output produced by a classifier is binary state of wet or dry in each missing site. To measure the performance of classifiers, we used the hit rate (H) which is an indicator of making correct wet and dry classifications among all trials (Wilks, 1995) as:

$$H = P_{w,\hat{w}} + P_{D,\hat{D}} \quad (6)$$

where  $P_{w,\hat{w}}$  is the probability of both observed and predicted occurrences are wet, and  $P_{D,\hat{D}}$  is the probability of both occurrences are dry. An asymptotic H value for a two-state pure random model is 0.5, while a maximum H value is 1.0 (a perfect model).

Table 2 summarizes the hit rates by classifiers at an arbitrary selected station (#13). It should be noted that the RND is a single site model while the others are a spatial model. The conventional Markov model (RND) performs slightly better (about 9 per cent) than a pure random model, as its abbreviation implied. However, the classification models improve significantly compared with those of the RND model. Especially, very promising results are shown when neural network based classifiers are used. At this station, the PNN with the Option 2 input features (x-y coordinates and model-driven stage) is superior to the other classifiers. The PNN has about 90 per cent of accuracy when it is used to classify rainfall occurrences. The H statistics for the other nine selected sites show the nearly identical improvements as observed at station #13.

Instead of presenting all performance statistics, Table 3 summarizes the best input feature option as well as classifier by site and by season. For example, at station #9, the ARB with the Option 3 input performs best on the wet season. In general, including model-driven stage to input feature (input Options 2 and 3) improves the performance of rainfall pattern classification. This general rule is not applicable to the site #44 which is located near the shore line so that stage is not influenced by the rainfall. For most stations, neural networks-based classifiers produce about 85% hit rates, which are quite promising. The result indicates that selecting input features plays an important role and improves rainfall modeling. It should be noted that the prediction of rainfall occurrence could be improved using more sophisticated input features such as, concurrent hydrologic measurements, radar data, cloud covers, soil moistures, or others. However, this additional task is out of scope here. Among the tested classifiers, the PNN is better for wet season prediction, while the LVQ outperforms for dry season prediction. Note that both the LVQ and the PNN have a competitive layer as a hidden layer to cluster input vectors in advance, which enhance the ability of the network for pattern classification.

## **6. Evaluation of Rainfall Amount Process**

Michaelides et al. (1995) and Kalogirou et al. (1997) examined the applicability of MNNs to estimate the amount of missing rainfalls in Cyprus. One of the issues on the MNNs is computing time. Rainfalls are randomly missing in time and space. Thus, we need to build a model for each day at each missing site, requesting tremendous time and effort though feasible with the current computing capability. For instance, the computation times (Pentium IV with 3.2 GHz) between a stepwise regression and a MNN model fitted to selective 100 missing data as a test run are 44 and 2319 seconds, respectively. There are a total of 9,400 (35%) rainy days out of

27,000 missing values from 63 stations. Expanding both approaches to all 9,400 missing values will take about 1.1 hours and 2.6 days, respectively. The stepwise regression is about 50 times faster than that of MNN at the same accuracy (the RMSE is 12.4 millimeters for both models). The effort that we have tried to make general training datasets for to save the computation time was unsuccessful since the data are missing at random in space and time. We concluded from this test that the MNN approaches are applicable for modeling rainfall occurrence but not very efficient for modeling rainfall depth.

The dimension of our regression model is quite large because there are a large number of rainfall gages (N=63). Or, there will be a multicollinearity problem from some redundant sites if all recorded sites are used as independent sites. To solve these problems, two regression models were tested: stepwise regression and principal components regression. Stepwise regression relies on an automated search procedure to select significant predictors (Hamilton, 1992). This approach uses stopping rules based on F-statistics or p-values to increase  $R^2$ . Principal components regression tries to model with reduced independent variables based on the information of a principal components (PCs) analysis. This approach selects and uses key component variables while accounting for patterns of variation among predictors. The advantage of using these two approaches is to make more parsimonious models without losing the accuracy of outcomes.

The two regression models were evaluated by the following skill score (SS) which is a percentage improvement of root mean squared error (RMSE) over that of a reference model as (Wilks, 1995):

$$SS(\%) = \frac{RMSE_m - RMSE_r}{RMSE_p - RMSE_r} \times 100 = \left(1 - \frac{RMSE_m}{RMSE_r}\right) \times 100 \quad (7)$$

where  $RMSE_p$ ,  $RMSE_r$ , and  $RMSE_m$  denote the RMSEs estimated with a perfect model, a

reference model, and an estimation model, respectively. The RMSE of an idealized perfect model ( $RMSE_p$ ) is zero. The reference model is the model to be served as a benchmark. The estimation model is the model constructed in this study.

The distribution of spatial daily non-zero rainfalls in South Florida is positively skewed. This may create a heteroscedasticity problem in modeling. Thus, the data were log-transformed. We also tested the Box-Cox transformations, but it was inferior to the log-transformation (the result is not shown here). To avoid the addition of stations having different geographical features, 10 nearest stations of the station of interest were selected in advance, and then regression analyses were performed. Table 4 shows the comparison of RMSEs for two different regression models. Overall, a stepwise regression outperforms the principal component regression. On average, the RMSE of both models are less than 1 per cent. Dry season errors are a half of wet season errors, but the relative errors will be opposite because the dry season rainfall is only 30% of the annual rainfalls. The errors in site #11 are higher than the other sites because it is isolated from the other sites.

Table 5 summarizes the SS statistics of different simulations using a stepwise regression model, where the reference model uses a sample mean from neighboring sites as the best estimate. The log transformation increase about 4% on average. The performance of model with classification is superior to that of model without classification. Although the performances of the proposed model are poor at stations #13 and #14, the proposed model improved skill scores up to 30% and 10% compared with a mean replacement and the without classification model, respectively. The proposed model would improve the accuracy of conventional methods up to 50% for estimating missing rainfall when a perfect classification is achieved.

In summary, the probabilistic neural network (PNN) approach is recommended to model

rainfall occurrence in South Florida because it needs shorter training time with satisfactory performance. The learning rule of the PNN is relatively simple. That is, it needs only a single pass through the training data (Duda et al., 2000). The stepwise regression with a logarithm transformation is also recommended to estimate the amount of missing rainfalls. It has been argued that one of shortcomings of a rainfall model with a regression estimator is that it underestimates the spatial variability of estimated rainfalls compared with that of observed ones. However, this problem was solved by removing the effect of zero values on dry days on the regression. With the proposed model, we could estimate the missing value at less than one per cent error.

## **7. Overall Performance Measures of the Proposed Model**

The previous two sections evaluate many alternative components of the proposed rainfall model. This section applied the model at all 63 stations to fill in actual gaps. The main focus in the model application here is to check whether the sample statistics are preserved without biases after filling in or not. Figure 6, for example, shows the comparisons of cumulative distributions of daily rainfalls in August at site #13. The raw data represent observed daily rainfalls on wet days and the filled-in data contains estimated daily rainfalls on missing wet days, which are not included in the raw data, using the proposed model. The missing rate on a given month is 31.8%, which is quite high. Although this case is one of extreme cases, there exist no significant differences between two distributions. The filled-in data passed the goodness-of-fit tests such as Chi-square and K-S test for all stations and all seasons (The test results are not shown here). Therefore the proposed model provides complete dataset for hydrologic analysis and modeling without distorting the distribution of data of interest.

Figures 7 through 8 show the changes of statistics before and after filling in. As similar to Figure 6, the raw data having observed daily rainfalls without missing values are used for the observed distributions, and the filled in data consisting of estimated daily rainfall on missing days using the proposed model are used for the estimated distributions. The results indicate that the proposed rainfall model has no significant bias on its estimates. Figure 7 is a typical rainfall occurrence problem where the wet-dry boundaries of both maps are the same, while a slight difference on the northern area is due to a technical interpolation problem. Missing rate on this date is 16 per cent. Figure 8 (a) is related to the rainfall occurrence. There is a spatial trend along the S-N direction: Low rainy days in the Florida Bay area, while high rainy days in the eastern coastal area. Figure 8 (b) is related to the rainfall amount. General trend along the N-S direction is preserved. The northern boundary has big changes, meaning that the proposed model should be careful for selecting the model boundary problem. It is safe to include more stations on the boundary if possible. In general, the differences of site statistics used in Figure 7 and 8 are dependent on the missing rate. However, this does not affect the distribution of primary statistics. The ANOVA test results (the test results are not shown here) indicate that the null hypothesis that the statistics from the observed (before filling in) and estimated (after filling in) distributions are same is not rejected.

## **8. Conclusions**

The main objective of this study is to develop a daily rainfall model that will be used specifically to fill in missing daily rainfalls. To fill in missing gaps, we could take advantage of concurrent measurements. In fact, a preliminary analysis showed that the spatial dependence is more significant than the serial dependence in daily rainfall and thus the model was proposed on

that way. The proposed model is based on a two step approach: The first relies on a pattern classification technique to determine the occurrence of rainfall on a missing day. Then, the second step uses an explicit regression model to estimate the amount of rainfall if the station is wet. Both steps use the spatial information as inputs.

A large portion of daily rainfall data is zero, which could introduce severe bias when the model is built in a straight way. The proposed model solved the intermittence by employing pattern classification as a prior step. A classifier evaluates the occurrence of missing stations using the spatial features of rainfall occurrence. Based on comparative studies performed in this study, a parsimonious model contributed to generalize the pattern classification for novel patterns, save the computation time, and provide sound estimates of the missing rainfall in the South Florida region. The result of validation revealed that the proposed model is robotic and unbiased.

In this study, the use of model-driven data such as relative water stages providing physical interaction with rainfall lead to good pattern recognition for rainfall occurrence. In recent years, with the advance in computer technology, promising researches for data gathering have been taking place to examine the possibility for using satellite and remote sensed data to monitor atmospheric activities on the continental or global scale (see Heymsfield et al., 1996, Levizzani et al., 2002, and National Research Council, 2003). These satellite-driven or remote sensed data might provide estimates of the rainfall occurrence as well as the rainfall intensity even in ungauged sites and at less than daily interval. With overcoming great computational expenses of the integration of the data from the high resolution dynamical models with the local-based indicator such as soil moisture, the more feasible and accurate imputation model could be developed.

The proposed rainfall model has many components and ramifications to apply, which include data transformation, rainfall occurrence, estimating rain depths, neural network solvers, etc. We carefully tested each of the components and presented the best solution for South Florida rainfalls. The recommended modeling procedure presented here will be valid for the sub-tropical regions where convective storms are dominant. However, the procedure could be changed for different areas or different rainfall regimes and the effect would be expected to improve the ability of explicit model for rainfall depth. Further testing of proposed approach to the different climatic conditions remains to the readers of this paper.

## **Acknowledgments**

This study was supported by the Everglades Research Fellowship Program in University of Florida. This program was funded by Everglades National Park. However, the views expressed in this article do not necessarily represent the views of the ENP/NPS.

## **References**

- ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, Artificial neural networks in hydrology. II: Hydrologic applications, *Journal of Hydrologic Engineering*, 5(2), 124-137, 2000.
- Batista, G., and M.C. Monard, An Analysis of Four Missing Data Treatment Methods for Supervised Learning, *Applied Artificial Intelligence*, 17(5-6), 519-533, 2003.
- Cooley, W.W., and P.R. Lohnes, *Multivariate Data Analysis*, 364 pp., John Wiley & Sons, Inc., New York, N.Y., 1971.
- Dempster, A.P., N.M Laird, and D.B. Rubin, Maximum likelihood estimation from incomplete

- data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39, 1-38, 1977.
- Demuth, H., and M. Beale, *Neural Network Toolbox: For Use with MATLAB, Version 4*, 846 pp., The Math Works, MA, 2000.
- Duda, R.O., P.E. Hart, and D.G. Stork, *Pattern Classification*, 654 pp., John Wiley & Sons, New York, N.Y., 2000.
- Dudoit, S., J. Fridlyand, and T.P. Speed, Comparison of discriminant methods for the classification of tumors using gene expression data, *Technical Report #576*, University of California, Berkeley, C.A., 2000.
- Haykin, S., *Neural Networks: A Comprehensive Foundation*, 696 pp., MacMillan, New York, N.Y., 1994.
- Hamilton, L.C., *Regression with Graphics; A Second Course in Applied Statistics*, 363 pp., Duxbury Press, Belmont, C.A., 1992.
- Hanson, C.L., K.A. Cumming, D.A. Woolhiser, and C.W. Richardson, *Microcomputer Program for Daily Weather Simulation in the Contiguous United States*, 38 pp., U.S. Department of Agriculture, Agricultural Research Service, ARS-114, 1994.
- Heymsfield, G.M., I.J. Caylor, J.M. Shepherd, W.S. Olson, S.W. Bidwell, W.C. Bonczyk, and S. Ameen, Structure of Florida thunderstorms using high-altitude aircraft radiometer and radar observations, *Journal of Applied Meteorology*, 35, 1736-1762, 1996.
- Kalogirou, S.A., C.C Neocleous, S.L. Michaelides, C.N. and Schizas, A time series reconstruction of precipitation records using artificial neural networks, *Proceedings of the EUFIT'97 Conference*, 2409-2413, 1995.
- Levizzani, V., R. Amorati, and F. Meneguzzo, *A Review of Satellite-Based Rainfall Estimation*

- methods*, 66 pp., European Commission Project MUSIC Report (EVK1-CT-2000-00058), Bologna, Italy, 2002.
- Maier, H.R., and G.C. Dandy, Neural networks for the prediction and forecasting of water resources variables; a review of modeling issues and applications, *Environmental Modelling and Software*, 15, 101-124, 2000.
- Michaelides, S.L., C.C. Neocleous, and C.N. Schizas, Artificial neural networks and multiple linear regression in estimating missing rainfall records, *Proceedings of the DSP'95 International Conference on Digital Signal Processing*, 668-673, 1995.
- National Research Council, *Satellite Observations of the Earth's Environment: Accelerating the Transition of Research to Operations*, 163 pp., The National Academies Press, Washington, D.C., 2003.
- Rajagopalan, B., U. Lall, and D.G. Tarboton, Nonhomogeneous Markov model for daily precipitation, *Journal of Hydrologic Engineering*, 1(1), 33-40, 1996.
- Roldán, J., and D.A. Woolhiser, Stochastic daily precipitation models: 1. A comparison of occurrence processes, *Water Resources Research*, 18(5), 1451-1459, 1982.
- Salas, J.D., Analysis and modeling of hydrologic time series, in *Handbook of Hydrology*, edited by D. R. Maidment, chap. 19, McGraw Hill, New York, N.Y., 1993.
- Schneider, T. (2001). "Analysis of incomplete climate data: Estimating of mean values and covariance matrices and imputation of missing values." *Journal of Climate*, 14, 853-871.
- SFWMD, *A Primer to the South Florida Water Management Model (Version 3.5)*, 233 pp., SFWMD, West Palm Beach, FL, 1999.
- Tarboton, K.C., C.J. Neidrauer, E.R. Santee, and J.C. Needle, Regional Hydrologic Modeling for Planning the Management of South Florida's Water Resources through 2050, ASAE

- Paper No. 992062, 1999.
- Wasantha Lal, A.M., Modification of canal flow due to stream-aquifer interaction, *Journal of Hydraulic Engineering*, 127(7), 567-576, 2001.
- Wilks, D.S., Multisite generation of a daily stochastic precipitation generation model, *Journal of Hydrology*, 210, 178-191, 1998.
- Wilks, D.S., *Statistical Methods in the Atmospheric Sciences An Introduction*, 467pp., Academic Press, San Diego, C.A., 1995.
- Winsberg, M., *Climate of Florida*, Florida State University, Tallahassee, FL, 2004.
- Woolhiser, D.A., Modeling daily precipitation – Progress and problems, in *Statistics in the Environmental and Earth Sciences*, edited by A.T. Walden and P. Guttorp, 71-89, E. Arnold, London, 1992.
- Woolhiser, D.A., and J. Roldán, Stochastic daily precipitation models: 2. A comparison of distributions of amounts, *Water Resources Research*, 18(5), 1461-1468, 1982.

Table 1. Summary statistics and missing rate at 10 selected stations used for model performance tests.

| Station ID | Station Name | Period of Record (Start year) * | Daily Max (mm) | Annual Mean (mm) | Missing Rate (%) |
|------------|--------------|---------------------------------|----------------|------------------|------------------|
| 7          | EVC          | 1949                            | 368            | 1389             | 4.5              |
| 9          | FLA          | 1962                            | 208            | 1211             | 0.1              |
| 10         | FMB          | 1949                            | 233            | 1345             | 0.3              |
| 11         | G54          | 1957                            | 267            | 1452             | 1.8              |
| 13         | IFS          | 1914                            | 213            | 1507             | 17.8             |
| 14         | JB           | 1991                            | 345            | 985              | 6.5              |
| 18         | MIAMIFS      | 1914                            | 310            | 1294             | 1.4              |
| 32         | P35          | 1982                            | 272            | 1389             | 13.0             |
| 41         | RPL          | 1949                            | 243            | 1466             | 2.1              |
| 44         | S20F         | 1968                            | 202            | 1210             | 1.1              |
| Average**  |              |                                 | 244            | 1385             | 10.6             |

\* The common ending date of records is December 31, 2003. The data of 1965-2000 were used for sample statistics.

\*\* Average values for all (63) stations in the study area.

Table 2. Hit rates at station #13 to measure performances of each classifier and input features to estimate rainfall occurrences.

| Station ID | Input Feature Option* | Conventional Model |       | Neural Network-Based Model |       |       |              |
|------------|-----------------------|--------------------|-------|----------------------------|-------|-------|--------------|
|            |                       | RND                | LDA   | LMB                        | ARB   | LVQ   | PNN          |
|            | 1                     | 0.585              | 0.720 | 0.750                      | 0.810 | 0.720 | 0.750        |
| 13         | 2                     | 0.585              | 0.695 | 0.820                      | 0.795 | 0.820 | <b>0.895</b> |
| (IFS)      | 3                     | 0.585              | 0.735 | 0.840                      | 0.760 | 0.780 | 0.875        |

Table 3. The best input option and classifier by site, which have been selected based on the hit rate that measures performance of rainfall occurrence estimates.

| Station<br>ID | Wet Season<br>(Jun.-Oct) |            | Dry Season<br>(Nov.-May) |            |
|---------------|--------------------------|------------|--------------------------|------------|
|               | Input Option             | Classifier | Input Option             | Classifier |
| 7             | 1, 2                     | LDA, PNN   | 2                        | LVQ        |
| 9             | 3                        | ARB        | 3                        | LVQ        |
| 10            | 3                        | PNN        | 3                        | LVQ        |
| 11            | 2                        | LDA        | 2                        | LDA        |
| 13            | 2                        | PNN        | 2                        | PNN        |
| 14            | 1                        | PNN        | 2                        | LVQ        |
| 18            | 3                        | LDA        | 2                        | ARB        |
| 32            | 2                        | PNN        | 2                        | LVQ        |
| 41            | 1                        | LMB, PNN   | 2                        | ARB        |
| 44            | 1                        | PNN        | 1                        | ARB        |

Table 4. Root mean square error (in mm) statistics of the estimated rainfalls using the proposed rainfall model, where S-Reg. is the stepwise regression and PC-Reg. is the principal component regression.

| Station ID | Wet Season |         | Dry Season |         | Yearly |         |
|------------|------------|---------|------------|---------|--------|---------|
|            | S-Reg.     | PC-Reg. | S-Reg.     | PC-Reg. | S-Reg. | PC-Reg. |
| 7          | 9.5        | 9.4     | 5.9        | 6.7     | 7.9    | 8.2     |
| 9          | 8.9        | 9.3     | 6.8        | 8.0     | 7.9    | 8.7     |
| 10         | 10.5       | 12.2    | 2.7        | 3.8     | 7.7    | 9.1     |
| 11         | 18.3       | 18.3    | 14.1       | 13.2    | 16.4   | 15.9    |
| 13         | 6.7        | 13.0    | 5.6        | 10.5    | 6.2    | 11.8    |
| 14         | 9.5        | 11.6    | 3.7        | 6.0     | 7.2    | 9.3     |
| 18         | 6.3        | 12.3    | 5.2        | 8.0     | 5.8    | 10.4    |
| 32         | 9.5        | 11.0    | 3.5        | 5.1     | 7.1    | 8.6     |
| 41         | 11.2       | 13.3    | 5.2        | 7.2     | 8.7    | 10.7    |
| 44         | 9.6        | 10.5    | 3.3        | 3.1     | 7.2    | 7.7     |
| Average    | 10.0       | 12.1    | 5.6        | 7.2     | 8.2    | 10.0    |

Table 5. Skill score (%) of models for estimating rainfall depths with a step-wise regression model.

| Station ID | Rainfall Depth Model (w/o transform) | Rainfall Depth Model (w/ log-transform) | Occurrence & Depth Models | Observed Occurrence & Depth Models |
|------------|--------------------------------------|---|---------------------------|------------------------------------|
| 7          | -0.9                                 | 8.6                                     | 9.7                       | 14.2                               |
| 9          | 0.9                                  | 10.0                                    | 11.2                      | 13.3                               |
| 10         | 12.4                                 | 14.4                                    | 16.8                      | 19.2                               |
| 11         | 2.7                                  | 3.7                                     | 5.0                       | 5.8                                |
| 13         | 31.4                                 | 28.0                                    | 29.2                      | 30.6                               |
| 14         | 25.2                                 | 23.6                                    | 23.8                      | 32.8                               |
| 18         | 42.8                                 | 43.4                                    | 44.0                      | 49.0                               |
| 32         | 7.8                                  | 12.8                                    | 14.3                      | 17.9                               |
| 41         | 7.4                                  | 13.5                                    | 15.7                      | 17.4                               |
| 44         | -3.0                                 | 5.3                                     | 6.9                       | 12.1                               |

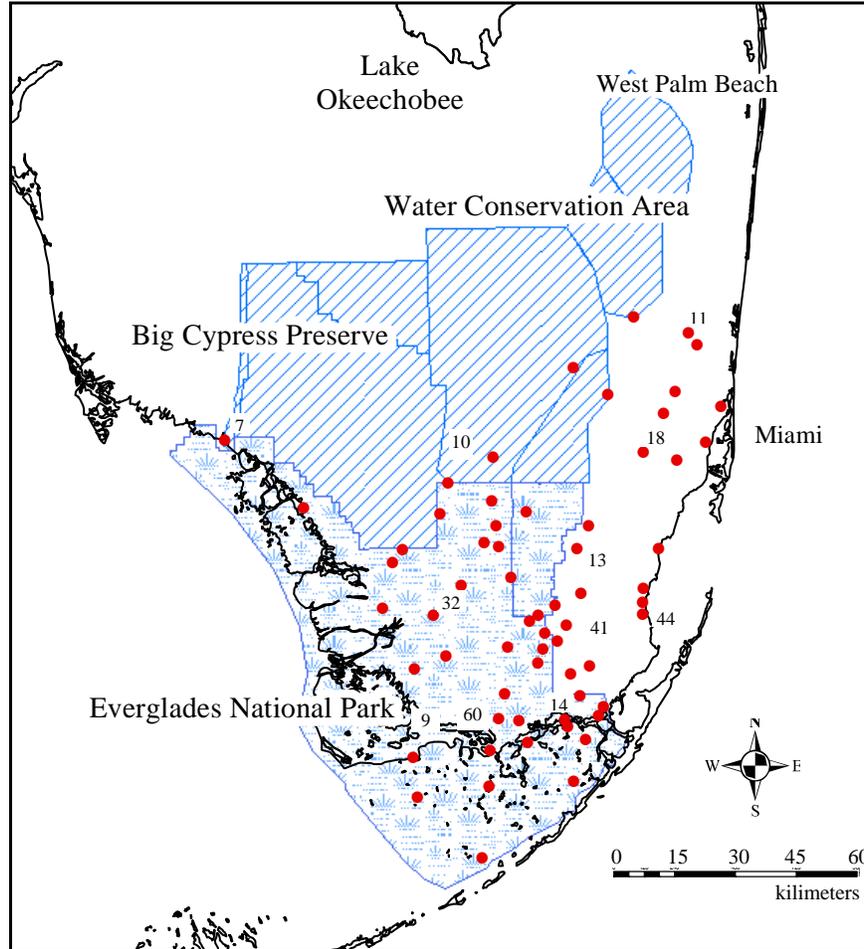


Figure 1. Rainfall stations (dots) in South Florida, where 10 sites (dot with ID number (7-44)) are selected for testing the proposed model.

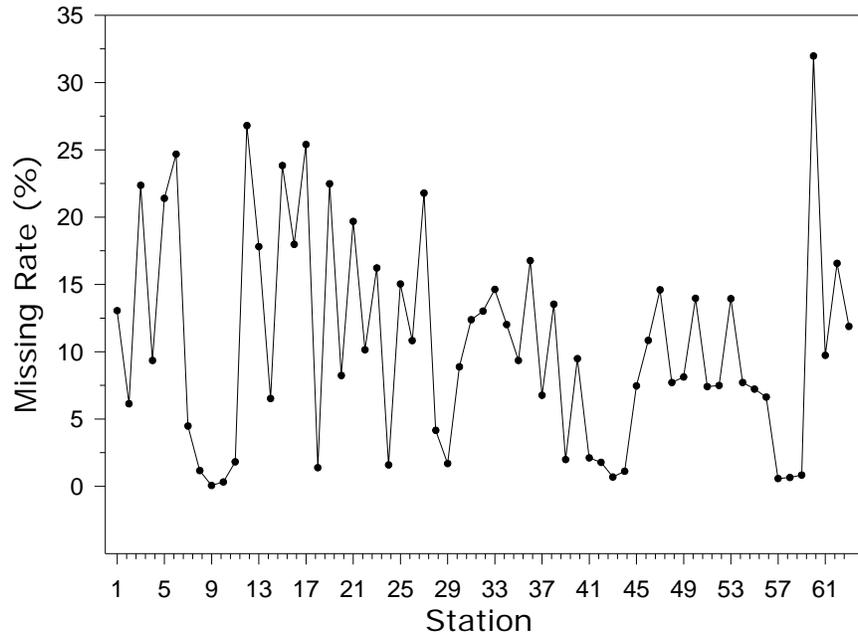


Figure 2. Missing rate at each station.

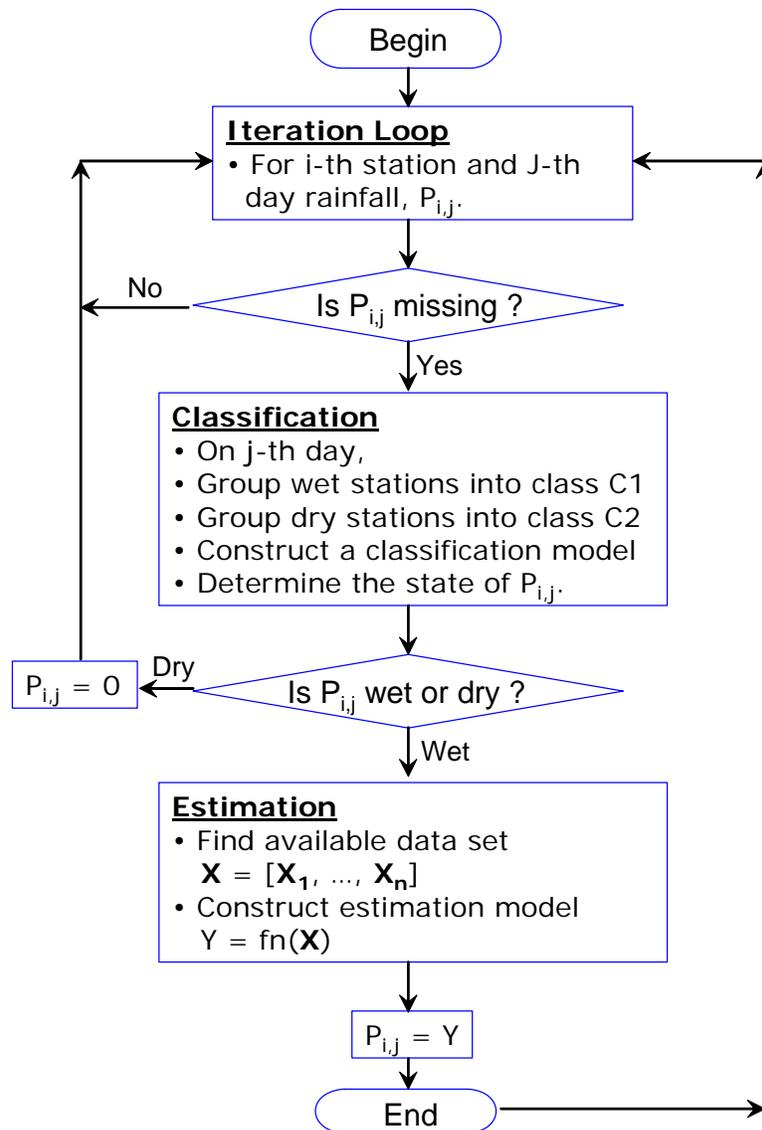


Figure 3. Structure of the proposed daily rainfall model for estimating missing gaps.

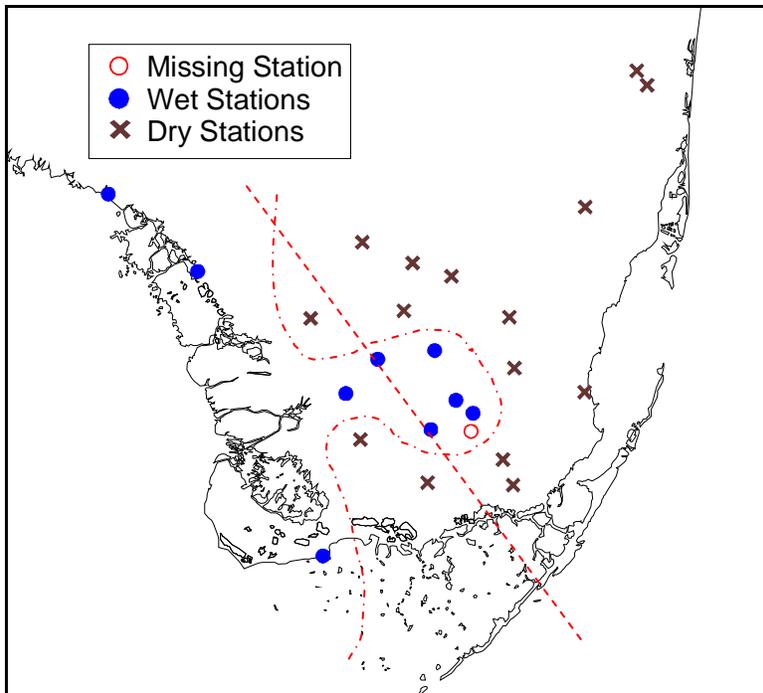


Figure 4. Sample rainfall occurrence boundary on June 9, 1990 on which 24 measurements available (dot and x marks). The straight line indicates an imaginary linear boundary to classify the rainy/no-rainy boundary, while the curved line is an imaginary nonlinear boundary based on measured data.

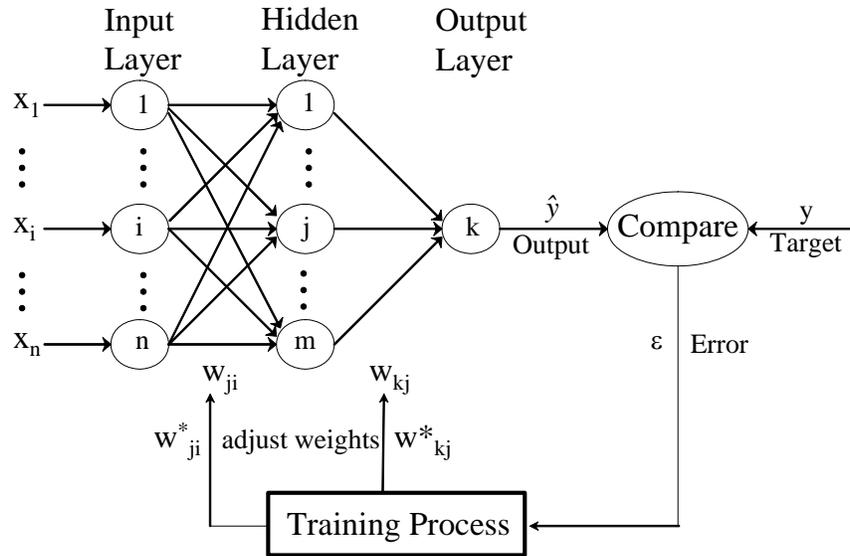


Figure 5. Typical schematic diagram of three-layer neural networks.

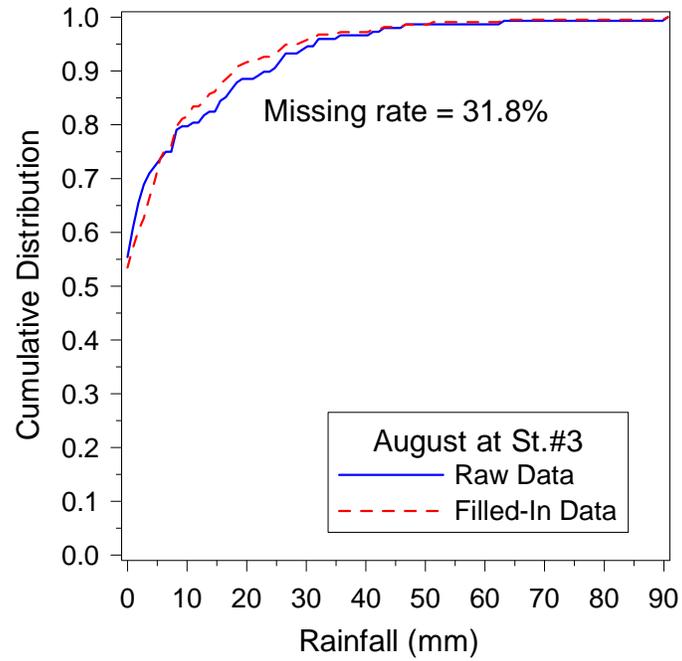
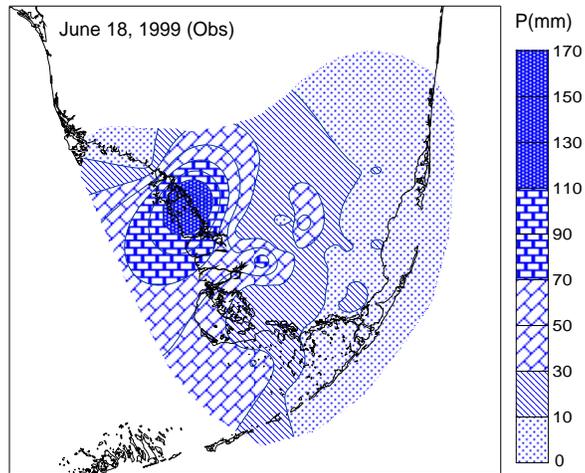
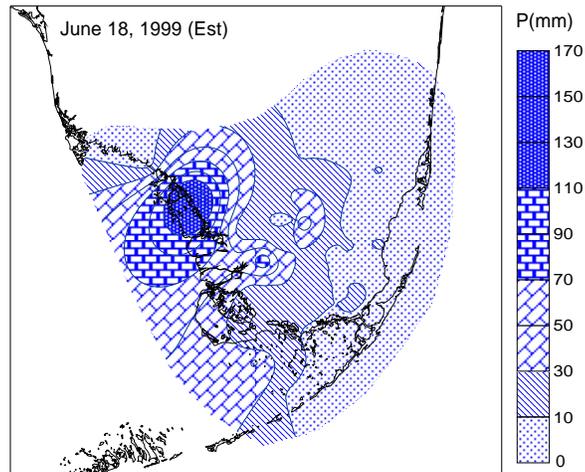


Figure 6. Comparison of cumulative distributions between raw data and filled-in data for August at station #13.

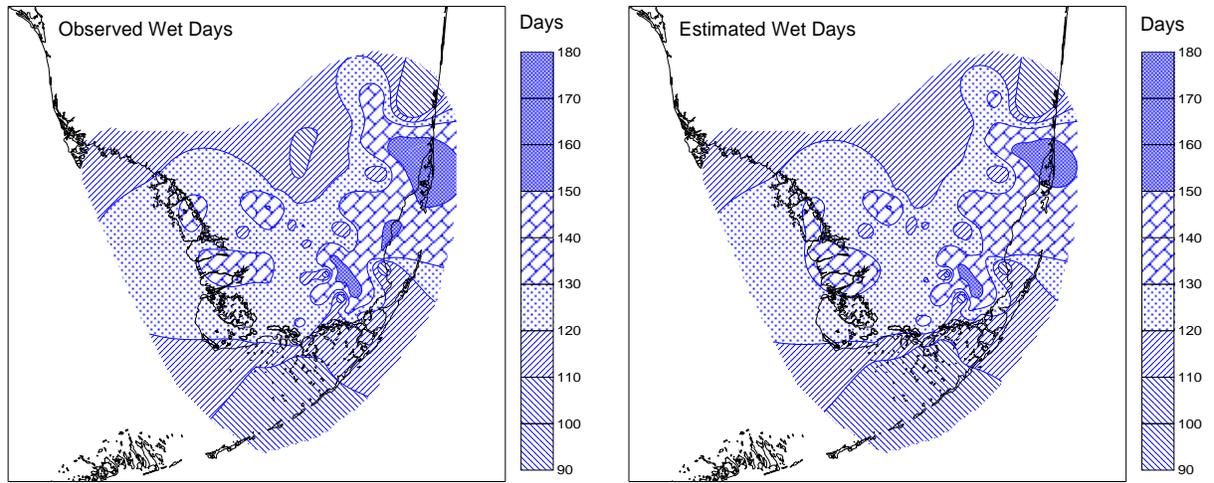


(a)

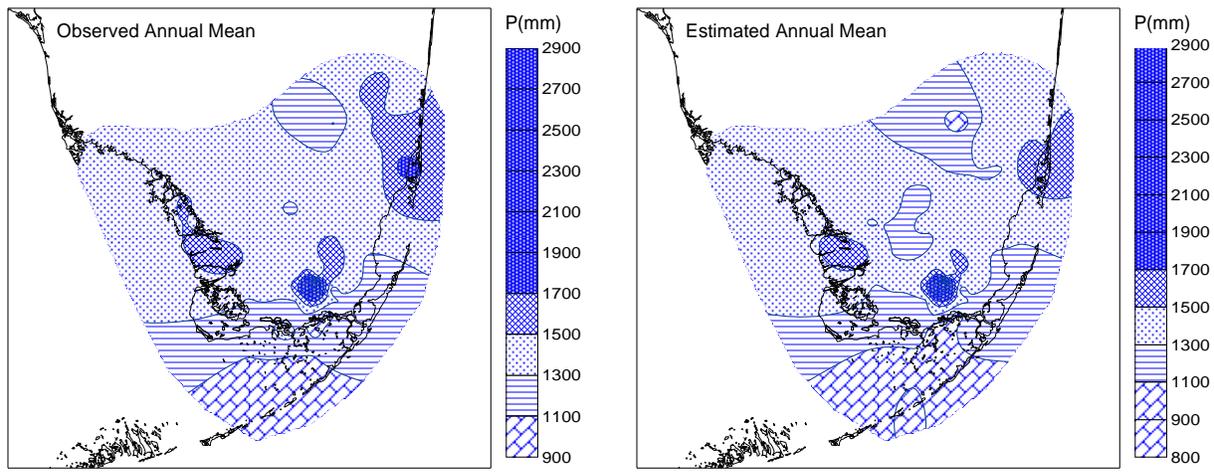


(b)

Figure 7. Rainfall distributions on June 18, 1999 (a) before (b) after filling in missing rainfalls.



(a)



(b)

Figure 8. Spatial distributions of a number of wet days (a) and annual mean rainfalls (b).